DOCUMENT RESUME

ED 292 876 TM 011 320

AUTHOR Oescher, Jeffrey
TITLE A Comparative Analysis of Three Item Selection Procedures.
PUB DATE Nov 87
NOTE 107p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (16th, Mobile, AL, November 11-13, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Addition; *Comparative Analysis; Grade 1; *Item Analysis; Primary Education; Subtraction; Test Construction; *Test Items
IDENTIFIERS Rasch Scaled Scores

ABSTRACT

This study empirically examined the proposition that different item selection procedures contribute to the development of achievement tests that optimally serve different purposes. A test of 110 addition and subtraction problems was given to a sample of 371 first-grade students. Three subsets of 30 items were selected from this 110-item test. The first included those items with the highest classical item discrimination indexes, the second of those best describing the continuum of Rasch item difficulties of the 110-item test, and the third of randomly selected items. The study examined which set of items best reflected student performance on the set of 110 items if the purpose of such testing was the differentiation of students, location of students on an achievement continuum, or generalizing student performance to the domain of 110 items. Primary analyses of the data indicated that the subtests consisting of classically discriminating items and those describing the continuum of Rasch difficulties yielded statistically different results from all other tests for differentiating students or generalizing performance to the domain of items. Secondary analyses indicated that these differences were attributed to the performance of the lower quartiles of students. (Sixteen tables are appended.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED292876

A Comparative Analysis of Three Item Selection Procedures

Jeffrey Oescher

Department of Educational Leadership and Foundations
University of New Orleans
New Orleans, LA 70148
504-286-6649

Paper presented at the annual meeting of the
Mid-South Educational Research Association
November 1987

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JEFFREY OESCHER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011320

BEST COPY AVAILABLE

"A Comparative Analysis of Three Item Selection Procedures"

:

Jeffrey Oescher, University of New Orleans

This study empirically examined the proposition that different item selection procedures contribute to the development of achievement tests that optimally serve different purposes. A test of 110 addition and subtraction problems was given to a sample of 371 first-grade students. Three subsets of 30 items were selected from this 110 item test. The first included those items with the highest classical item discrimination indexes, the second of those best describing the continuum of Rasch item difficulties of the 110 item test, and the third of randomly selected items. The study examined which set of items best reflected student performance on the set of 110 items if the purpose of such testing was the differentiation of students, location of students on an achievement continuum, or generalizing student performance to the domain of 110 items.

Primary analyses of the data indicated the subtests consisting of classically discriminating items and those describing the continuum of Rasch difficulties yielded statistically different results from all other tests for differentiating students or generalizing performance to the domain of items. Secondary analyses indicated these differences were attributed to the performance of the lower quartiles of students.

TABLE OF CONTENTS

	Page
CHAPTER ONE	
Nature and Significance of the Problem	1
CHAPTER TWO	
Review of the Literature	7
CHAPTER THREE	
Methodology	18
CHAPTER FOUR	
Results	34
CHAPTER FIVE	
Summary, Conclusions, and Implications	63
REFERENCE NOTES	71
REFERENCES	72
FOOTNOTES	75
APPENDICES	77

LIST OF TABLES

		Page
Table 1:	Summary of Validity Studies	14
Table 2:	Summary of Reliability Studies	15
Table 3:	Summary Item Statistics - Classical Difficulties	36
Table 4:	Summary Item Statistics - Classical Discrimination Indexes	36
Table 5:	Summary Item Statistics - Rasch Difficulties	36
Table 6:	Summary Test Statistics - Raw Scores	39
Table 7:	Summary Test Statistics - Rasch Abilities	40
Table 8:	Items Included on the CT, LTT, and ST	44
Table 9:	Descriptions of Rasch Item Difficulty Continuums for the CT, LTT, and ST	48
Table 10:	Raw Score Correlations Between Tests	53
Table 11:	Repeated Measures ANOVA for Rasch Abilities	57
Table 12:	Repeated Measures ANOVA for Proportions of Correct Responses	57
Table 13:	A Summary of the Comparisons of the UT with the CT, LTT, and ST	59
Table 14:	A Summary of the Results of the Secondary Analysis	62

LIST OF FIGURES

Page

Figure 1: Distribution of Raw Scores on the UT 42

Chapter One

Nature and Significance of the Problem

Introduction

In recent years the use of test information in schools has increased dramatically. With this increase, a number of issues concerning the development, administration, interpretation, and use of tests and test results have arisen. These issues are no longer restricted to discussions among measurement experts; they have become highly public and intensely political. The public has become keenly aware of such phrases as "test bias," "norm-referenced," "minimum competency," and "truth in testing." State legislatures are debating testing issues and ratifying testing laws. State and federal courts are adjudicating testing controversies (Haney, 1980).

As testing has become more and more a public concern, the demand for theoretical and technical development in the measurement field has increased. Although a phenomenal amount of progress has been made over the last decade, many areas have yet to be thoroughly investigated. One of these is the use of item analysis data in the development of achievement tests designed to serve different purposes.

Tests of educational achievement can be used to differentiate among students, locate students on a continuum of knowledge, or generalize student performance from a small set of items (that is, a test) to a universe of items. It seems reasonable that different types of tests are best suited for these different purposes. To differentiate

among students, for example, tests need to be designed so as to distribute the population of students in terms of their relative performance. Items on such tests should accurately discriminate between high and low achieving students. To locate students on a continuum of knowledge, on the other hand, tests must permit the measurement of achievement on a well-defined, ordered, content domain. Correct answers to items on these tests must require different amounts or levels of achievement of the content domain. Finally, generalizing to a universe of instructionally relevant items requires that the tests be designed to adequately represent the universe. Items on these tests should all contribute to the validity of the inferences made from the sample of items to the universe.

The usefulness of an achievement test must be judged primarily on the basis of its validity. Ultimately, this validity is concerned with both the content underlying the test and the purpose for which the test is to be used. It is necessary, for example, to include on an achievement test only those items which sample the material included in the curriculum or taught in the course. The degree to which such sampling is acceptable is a measure of the test's content validity.

Although content validity is necessary, it is not sufficient to ensure the validity of an achievement test. Consideration must also be given to the purpose for which the test is designed. For example, a test must have acceptable variability if its purpose is to differentiate among students.

The purpose of this study is to empirically examine the proposition that different methods for selecting items contribute to the development

of achievement tests that optimally differentiate among students, locate students on a continuum of knowledge, or generalize student performance to an item universe. While an examination of recent measurement literature has identified extensive theoretical support for this proposition (Cook & Hambleton, 1978; Cox, 1966; Denham, 1975; Ebel, 1979; Gronlund, 1981; Popham, 1981; Sax, 1981; Subkoviak & Baker, 1978), substantive empirical support is almost nonexistant.

Problem

Three tests will be developed by selecting subsets of items from a large, well-defined universe of items. The first test will consist of items selected on the basis of classical item statistics, especially item discrimination. The second test will include items selected on the basis of one-parameter, latent-trait item statistics, especially item fit and difficulty. The third test will consist of items randomly selected from the universe of items. Considering the performance of students on these three tests as well as a test of the entire universe of items, the study will address the following questions:

1. When the purpose of testing is the differentiation of students, how do the results of the tests compare with one another?
2. When the purpose of testing is the location of students on a continuum of knowledge, how do the results of the tests compare with one another?
3. When the purpose of testing is the generalization of student performance to an item universe, how do the results of the tests compare with one another?

Significance of the Study

Testing in the public schools has become an increasingly important issue. In the last several years alone, 38 states have mandated minimum competency testing programs, and "truth in testing" bills have been introduced in 20 state legislatures. At least two federal court decisions (Debra P. v. Turlington and Larry P. v. Riles) have been rendered on testing issues. National debates have focused on the merits of standardized tests and testing services, and parents have become increasingly critical of the extensive amount of testing in their children's schools (Haney, 1980; Nairn & others, 1980; Brandt, 1980).

Educators have found themselves intimately concerned with these testing issues, because tests represent one of the best means by which objective information on student achievement can be obtained. This information, in turn, plays an important role in the selection, placement, and certification of students as well as the diagnosis of student and curricular problems.

The accuracy of these decisions is based at least partly on the accurate assessment of student achievement. This latter accuracy, defined as validity in measurement terminology, is contingent upon consideration of both test content and purpose. Typically, items are written to achieve content validity, whereas they can be selected so as to enhance the likelihood of accomplishing a specific purpose.

The existence and use of varying item analysis procedures implies that there is a possibility of differentially valid achievement tests, that is, tests that are more or less valid for different purposes, curricular, or students. This study will provide a greater understanding of the relationship between item selection procedures and the

development of tests optimally useful for accomplishing different purposes. Such an understanding would contribute to the development of tests with greater validity, which, in turn, would contribute to the precision of educators' decisions concerning student achievement.

Delimitations

1. This study will not examine achievement test content other than mathematics. This content area is unique in two respects. First, in some instances, it is possible to define a set of mathematics items which describe all of the potential exercises examining a specific instructional unit. Selecting test items from such a set ensures the development of tests with consistent degrees of content validity. Second, it is possible to objectively score math items. There is, for example, only one correct answer to the problem 4 + 2 = ___. Such objectivity reduces the possibility of confounding effects resulting from the inconsistent scoring of items.

As a content area, then, mathematics provides an opportunity to define an entire universe of objective items covering a specified instructional unit. These characteristics are essential to this study.

2. This study will not use a latent-trait model other than Rasch. The Rasch model provides a simple description of what happens when any person encounters any item. It states that the outcome will be entirely governed by the difference between the ability of the person and the difficulty of the item. The more able the person, the better their chances for success with any item. The easier the item, the more likely any person is to solve it (Wright & Stone, 1979).

Chapter Two

Review of the Literature

Introduction

An examination of recent measurement texts such as Popham (1981) and Sax (1981) indicated a great deal of theoretical support for differentially valid item selection procedures. Additionally, a computer search of the ERIC and Psychological Abstract Indexes from 1965 to the present identified several studies that provided further theoretical as well as some empirical support. Although a few relevant studies were found in reputable journals, much of the information contained in this review was found in AERA or NCME paper presentations and dissertation abstracts.

Although many studies were considered in this review, only those relevant to the comparison of item selection procedures have been included in the following discussion. Likewise, no attempt was made to discuss the theoretical foundations of classical, latent-trait, or domain-referenced test theory. This information has been well documented in the works of Gullikson (1950), Rasch (1980), and Hively (1974).

This review revealed two categories of studies relevant to the present study. The first consisted of generalized theoretical support, the second of validity and reliability investigations. Both of these categories are discussed below.

General Theoretical Support

Generalized theoretical support of differentially valid item selection procedures for different test purposes has been indicated by Cook and Hambleton (1978), Denham (1975), Subkoviak and Baker (1978), and recent measurement texts. Cook and Hambleton differentiated classical and latent-trait procedures relative to their use in the development of norm-referenced and criterion-referenced tests. They assert that since most norm-referenced tests are designed to differentiate among examinees, these tests should be designed to maximize discrimination among examinees. They imply that such test items should be selected on the basis of their item discrimination and item difficulty indexes. They note, however, that these item characteristics vary from sample to sample. Such sample dependency is quite consistent with the concept of differentiating students on the basis of their relative achievement.

Cook and Hambleton suggest that latent-trait theory affords a more powerful item selection method for criterion-referenced tests. Using this approach, items can be selected on the quality of information they contribute toward identifying a given ability level. Item difficulties are hypothesized to be sample invariant, that is, the estimations of these difficulties are free from the particular abilities of the persons in the calibration sample. As a result, items are calibrated in a way that can contribute to the measurement of examinee performance.

Similarly, Denham (1975) proposes a paradigm for a conceptual distinction between criterion-referenced tests (CRTs), domain-referenced tests (DRTs), and norm-referenced tests (NRTs). According to Denham, CRTs compare examinees to some criterion, DRTs sample items from a

specified set of tasks, and NRTs compare examinees to some specified norm group. Denham suggests that items for the CRTs, DRTs, and NRTs be selected by latent-trait, random sampling, and classical procedures respectively.

Subkoviak and Baker (1978) discuss general approaches to selecting items for tests which differentiate among students and those which are representative of a well-defined domain. They suggest that a test in which items are selected according to classical discrimination indexes should differentiate among students more accurately than one in which items are selected by other methods. Similarly, selecting items on statistically sound sampling procedures should develop a test which accurately reflects student performance on the entire domain of items.

Additional theoretical support was evident from an examination of recent measurement texts (Ebel, 1979; Gronlund, 1981; Popham, 1981; Sax, 1981). In their discussions relevant to test development, these authors imply that the unique contributions of varying item selection procedures enhance the development of different types of tests.

Validity and Reliability Considerations

A second category of relevant studies consisted of those that addressed the validity and reliability of various types of tests. Goh (1979), Crehan (1974), and Smith (1978) consider the efficiency of item selection procedures relative to the classification of examiness. Benson (1978), Benson, Crocker, & Ware (1978), Benson & Wilson (1979), and Hatton (1980) discuss the precision and efficiency of various item selection procedures.

Goh's (1979) study examined empirical and randomized item selection procedures relative to the creation of a short form of the Wechsler Intelligence Scale for Children-Revised (WISC-R). Two short forms were created. The first used an empirical procedure that considered classical difficulty, discrimination, and reliability indexes, while the second used a randomized selection process. The results of both tests were compared to the results of the WISC-R.

The results of Goh's analyses indicated that the empirical method yielded more accurate results than the randomized one. Both methods produced tests that showed statistically significant correlations with the scores on the original battery. However, the empirical short form produced statistically more accurate estimations of the mean IQs for the subjects as a group or as subgroups across different intellectual levels. It also correctly classified a higher percentage of the examinees relative to the cutoff scores for the various levels of intelligence reported in the WISC-R manual. Neither method, however, was judged to produce a test suitable for diagnostic/classification purposes.

Crehan (1974) compared six item selection methods (i.e., point biserial, random, Cox-Vargas, modified Brennan, post-test agreement, and teacher ranking) relative to the validity and reliability of teacher-made mastery tests. Validity was defined as the accuracy with which a test classified each student as a master or nonmaster of the behaviors measured. Validity indexes were computed as the proportion of correct classifications in a 2 x 2 matrix based on pass/fail and instructed/uninstructed criteria. Crehan assumed that students exposed to relevant instruction would be masters, whereas unexposed ones would not. Correct

classifications would be represented by students who either received instruction and passed the test or those who neither received instruction nor passed.

Reliability was defined as the consistency of individual mastery decisions across parallel test forms. Reliability indexes were calculated as were the validity indexes, except the 2 x 2 matrix was based on pass/fail criteria for each test form.

The resulting estimates of test validity and reliability were compared across item selection procedures. Crehan's analyses indicated that two selection methods, the Brennan and Cox-Vargas, produced higher validity indexes while none substantively affected the reliability indexes. The Brennan method utilized the difference between proportions of mastery and nonmastery examinees answering an item correctly, and the Cox-Vargas method used the differences between pre and post proportions of examinees answering an item correctly. Furthermore, Crehan indicated that the benefits of these two methods would not warrant the extra effort required on the part of a classroom teacher.

Smith (1978) examined the effects of five item selection methods (i.e., point biserial, random, Cox-Vargas, Brennan, and phi coefficient) on the classification accuracy and consistency of mastery tests. A test of 70 items served as an item pool. Five groups of 40 items were selected according to the criteria for each method. Next, two "tests" of 20 randomly selected items were constructed from each group of 40 items. An examinee's true state of mastery was determined by their performance on the original 70 item test. Accuracy was defined as the proportion of examinees correctly classified into their true mastery

state by the various 20 item tests. Consistency was examined through the replication of correct classification between the 20 item test forms.

An interesting aspect of Smith's study concerned the simulation of data using a two parameter latent-trait analysis of the 70 item test. Item characteristics were held constant, while examinee ability was varied to reflect varying degrees of instructional effectiveness. Each of the resulting data sets were then analyzed regarding both accuracy and consistency.

The results indicated that the most accurate and consistent methods varied according to the instructional effectiveness and the resultant distribution of scores. Under conventional instructional conditions in which a large amount of variability was present in the scores, the point biserial method was superior. In mastery learning situations where little variability is present in the score distribution, the Brennan and Phi coefficient methods were superior.

Benson (1978) compared three item selection methods (i.e., point biserial, factor analytic, and Rasch) relative to the efficiency and precision of a test. Efficiency was defined as the degree to which a test was able to detect small differences in examinees' abilities at specified levels. The point biserial and factor analytic methods were superior for very low and very high ability examinees, while the Rasch method was more appropriate for those students with average and high abilities.

Measures of precision included internal consistency and standard error of measure. The results indicated no significant differences among

methods regarding precision.

A similar study by Benson, Crocker, and Wade (1978) compared the same item selection methods relative to precision and efficiency. Similar findings were reported in this study.

Benson and Wilson (1979) conducted a study that differed from that of Benson, Crocker, and Wade (1978) only in the determination of test length. Their results indicated the Rasch method produced the most difficult and least internally consistent test. However, all observed differences were small. All three item selection methods were quite similar over most of the score range.

Hatton (1980) also replicated Benson's (1977) study and found that the classically developed tests differed significantly from the Rasch test only on item discrimination. Hatton concluded that due to the overall lack of any substantive empirical distinctions between the types of item selection methods, the sample invariance characteristics of the Rasch method afforded a superior selection method.

Conclusion

This review of the literature revealed several theoretical and empirical studies comparing the relative merits of item selection methods. These studies support the contention that differentially valid, precise, and efficient item selection procedures exist.

The empirical research examined tends to support the use of different item selection procedures. All but one of the studies found significant differences among various selection procedures with respect to test validity (see Table 1), even though few found differences relative to reliability (see Table 2).

Table 1

Summary of Validity Studies

Study	Purpose of Test	Operational Definitions	Item Selection Methods	Findings
Goh	IQ	Mean scores; Correct intellectual classification	Classical, random	Classical superior to random
Crehan	Mastery	Correct mastery classification	Point biserial, random, Cox-Vargas, modified Brennan, post-test agreement, teacher ranking	Modified Brennan and Cox-Vargas superior
Smith	Mastery	Correct mastery classification	Point biserial, random, Cox-Vargas, Brennan, Phi-coefficient	Varied effectiveness according to distribution of scores
Benson	Unspecified	Detection of small differences among examinees	Point biserial, factor analytic, Rasch	Varied effectiveness according to distribution of scores
Benson, Crocker, & Wade	Unspecified	Detection of small differences among examinees	Point biserial, factor analytic, Rasch	Varied effectiveness according to distribution of scores
Benson & Wilson	Unspecified	Detection of small differences among examinees	Point biserial, factor analytic, Rasch	None
Hatton	Unspecified	Detection of small differences among examinees	Point biserial, factor analytic, Rasch	Differences between point biserial and Rasch on item discrimination

Table 2

Summary of Reliability Studies

Study	Purpose of Test	Operational Definitions	Item Selection Methods	Findings
Crehan	Mastery	Replication of classification across forms	Classical, random	No differences
Smith	Mastery	Replication of classification across forms	Point biserial, random Cox-Vargas, modified Brennan, post-test agreement, teacher ranking	Varied according to underlying distribution of scores
Benson	Unspecified	Internal consistency; SEM	Point biserial, factor analytic, Rasch	No differences
Benson, Crocker, & Wade	Unspecified	Internal consistency; SEM	Point biserial, factor analytic, Rasch	No differences
Benson & Wilson	Unspecified	Internal consistency; SEM	Point biserial, factor analytic, Rasch	No differences
Hatton	Unspecified	Internal consistency; SEM	Point biserial, factor analytic, Rasch	No differences

Goh's study focused on a comparison of classical and random item selection procedures in the development of a short form of the WISC-R. His findings supported the use of classical, not random, procedures.

Crehan's and Smith's studies investigated the relative merits of item selection procedures for mastery tests. Due to the nature of mastery tests, validity decisions concern the accuracy with which students are classified as passing or failing.

Crehan's validity index was calculated with respect to a 2 x 2 matrix of pass/fail and instructed/uninstructed classifications. An assumption is made that instructed students would be considered masters, whereas uninstructed ones would not. Such an assumption should be viewed with a great deal of skepticism, especially when it is the basis for determining test validity. Nevertheless, the results once again indicated differentially valid item selection procedures.

Smith considered test validity relative to a more acceptable mastery classification. He also examined the classification issue relative to several samples of differing variability. The results again indicated test accuracy could be enhanced by the choice of a specific procedure for selecting items.

The four studies associated with either Benson or Hatton focused on the efficiency with which a test identified small differences in examinees' abilities at specified levels. Two of these studies indicated that different selection procedures were more efficient for different ability levels.

In conclusion, the existing empirical evidence indicates that the development of tests designed to either correctly identify masters and

non-masters or detect small differences in student abilities can be enhanced by the use of a specific item selection procedure. Unfortunately, similar evidence does not exist as to whether item selection procedures permit the optimal construction of tests for three purposes: the differentiation of students, measurement of students on a continuum of knowledge, or generalization of test results to a domain of items. Given the amount of theoretical consideration that has been focused on this issue, this lack of evidence is surprising.

Chapter Three

METHODOLOGY

Introduction

Chapter 1 has provided some insight into the nature and significance of the proposition that different item selection procedures contribute to the development of achievement tests that optimally serve different purposes. Three test purposes were identified: differentiating among students, locating students on a continuum of knowledge, and generalizing student performance from a small set of items to a universe of items. Three ways of selecting items for inclusion on a test were also identified. These procedures focused on classical item discrimination, latent-trait item fit and difficulty, and random sampling. The study proposed to examine the relationships between these test purposes and item selection procedures.

Chapter 2 identified consistent theoretical support for the problem addressed in the study. However, little empirical support was found in the literature for the particular problem underlying this study. Due to the lack of such support and the proliferation of papers espousing the merits of specific procedures, the present study was deemed justifiable.

Chapter 3 explains the specific procedures that were used to answer the research questions posed. The chapter is divided into separate sections which describe the sample, tests, test administration, statistical hypotheses, and data analysis.

Overview of the Study

A test of all possible single-digit addition and subtraction problems was administered to a sample of first grade students. The results of this test provided the information needed to develop three additional, shorter tests. The first of these consisted of a subset of 30 items selected on the basis of classical item statistics, especially item discrimination. The second consisted of a subset of 30 items based on the results of one-parameter, latent trait item statistics, especially the extent to which the item response data fit the hypothetical model and the difficulty of the items. The final test consisted of a subset of 30 items randomly selected from all of the items on the initial test. The results from all four tests were then compared with respect to the differentiation among students, the location of students on a continuum of knowledge, and the generalization of student performance to an item universe.

Each of the following sections provides the pertinent information regarding these comparisons. To avoid needless descriptive repetition, the test of all possible single-digit addition and subtraction problems will be referred to as the Universe Test (UT). The tests in which item selection was based on classical item statistics, latent-trait item statistics, and random sampling will be referred to as the Classical Test (CT), Latent-Trait Test (LTT), and Sampled Test (ST) respectively.

Population and Sample

The choice of a population for this study reflected the need for variability among student scores on the UT. Substantial variability would contribute to an accurate separation of students as well as a

clearer definition of a continuum of achievement, while slight variability would not. A sufficient amount of variation could be obtained from a population of students who had received instruction, but had not mastered, the addition and subtraction of single-digit numbers.

The addition and subtraction of single-digit numbers are objectives of first grade instruction. Several first grade teachers indicated that the majority of their students would be able to demonstrate mastery of the addition items after approximately 2-3 months of instruction, while mastery of the subtraction items would probably take several additional months. Most of the teachers agreed that a sufficient amount of variability would be present in the scores of the UT if the test was administered after 3-4 months of instruction. Consequently, the population was narrowed to those students who had been in the first grade for approximately 3-4 months.

The size of the sample to be drawn from this population was dependent upon the item analysis procedure requiring the most subjects. The latent-trait analysis was the most confining in this respect, requiring a minimum of 3-4 subjects per item (Ryan, Note 1). Since there were 110 items on the UT and all other tests were subsets of it, an attempt was made to sample approximately 440 students.

The actual initial sample consisted of 427 first grade students. These students were selected from four elementary schools in the Columbia, South Carolina area. The schools were representative of various socio-economic levels in order to insure a range of students in terms of math ability or achievement. Due to several administrative constraints imposed on this research, further description of the sample

is not possible. These constraints are discussed in the section of this chapter pertaining to the administration of the test.

The size of this initial sample was reduced for two reasons. First, 33 students failed to complete the UT. Although there were a few exceptions, absences during the administration of the test were the major reasons for these incomplete tests. Second, the scores of 23 students were judged to be invalid. Nineteen of these students came from a single class in which every subtraction problem was answered correctly by each student. The pattern of responses to the corresponding addition problems indicated that assistance was probably given to these students on the subtraction section. The additional 4 students were eliminated because their teacher indicated that they had not received any addition or subtraction instruction. Therefore, the final sample consisted of 371 students.

Development of the Universe Test

The UT consisted of 110 items. These items were generated by considering all possible combinations of two single-digit numbers under the condition that if the numbers in each item were unequal, the smaller number would be added to or subtracted from the larger one. Consequently, 55 addition and 55 subtraction items were prepared.

Each item was written in the following format:

$$\begin{array}{r} A \\ \underline{\text{OPERATION } B} \end{array}$$

where OPERATION was either addition (+) or subtraction (-),
A represents the larger of two single-digit numbers, and
B represents the smaller of two single-digit numbers.

Each item required students to record the answer immediately below the problem. All of the addition problems were presented on pages 1 and 2 of the test, while pages 3 and 4 contained only subtraction problems. The ordering of the items on the addition and subtraction sections was determined by a random process without replacement. A copy of the UT can be found in Appendix A.

A correct answer was the exact result of the required operation on the specified numbers; otherwise an answer was incorrect. The students received one (1) point for each correct answer and zero (0) points for each incorrect one. The total score for each student was the sum of the correct answers.

The content validity of the UT was assured by the inclusion of a defined set of items that reflected the classroom instruction on the topic. Objectivity was established by the use of a standardized scoring procedure. The reliability of the test was estimated by its standard error of measure (SEM). The SEM is considered less sensitive than other reliability estimates to differences in the variability of score distributions (Ebel, 1979, pp. 284-287; Gronlund, 1981; pp. 101-104; Popham, 1981, pp. 146-148; Sax, 1981, pp. 273-277).

Strategies for Constructing Subtests

Test Length. A review of the literature indicated several issues with respect to test length. The first was applicable only when specific decisions concerning student selection, placement, certification, or diagnosis were made from the test results (Millman, 1973; Huynh & Saunders, 1980, Note 2). Since there were no decisions implicit in the use of this test, this concern was alleviated.

A second issue involved the item parameters estimated in the latent-trait analysis. The BICAL-Version 3 computer program of Mead, Wright, and Bell (Note 3), used for the latent-trait analyses in this study, utilizes the unconditional (UCON) procedure for estimating item difficulty. According to Wright and Douglas (1977), the use of this procedure results in a biased estimate of item difficulty. Their research has indicated that the extent of bias is negligible when test length approaches 30 or more items.

Several other concerns were considered in the choice of a length for the tests. First, extremes needed to be avoided. A test of 105 items would be highly representative of the UT; a test of 5 items would not. Second, the length of all three tests (i.e., CT, LTT, and ST) needed to be equivalent to avoid confounding the comparisons of their scores and related validities. Finally, latent-trait item statistics needed to be examined prior to determining a test length (Benson & Wilson, 1979). For example, a 55 item test would be inappropriate if only 35 items fit the Rasch model.

Considering all of these factors, a test length of 30 items was chosen. An additional benefit of this choice was the similarity between this test length and that of teacher-made and statewide competency tests. This similarity could enhance the use of the results of the study.

Classical Test. The 30 items selected for the CT were those that best met classical item analysis criteria, especially item discrimination. Classical item analysis provides information by which items can be selected on the basis of their ability to differentiate among

students. This information is obtained from an examination of an item's classical discrimination index.

This index estimates the extent to which the item differentiates among students with high and low achievement (Gronlund, 1981, pp. 256-264; Sax, 1981, pp. 189-192). It is most frequently calculated by correlating the correctness of each student's response to each item with each student's total test score. The correlation is positive and high when those responding correctly to an item tended to do well on the test as a whole. It is negative when those responding correctly to an item tended to have the lowest scores.

It is theorized that items with discrimination indexes of +1.00 are perfect choices for inclusion on a test designed to differentiate students. Although obtaining an index of +1.00 is unrealistic, the index for certain items can approach this value. The closer to +1.00, the better the item differentiates among students.

Ebel (1979, p. 267) suggested that the minimally acceptable value for an item discrimination index was .40. Consequently, the items selected for the CT were those with indexes closest to 1.00, yet above .40. The TSSA2 computer program (Wolf & Klopfer, Note 4) was used to obtain all of the classical item analysis statistics.

Student responses to the 30 items identified for the CT were culled from their responses on the UT. Scores on the CT were calculated by summing each student's correct responses to this subset of items.

Latent-Trait Test. The 30 items selected for the LTT were those meeting one-parameter, latent-trait analysis (i.e., Rasch) criteria. This analysis provides information by which items can be selected on

the basis of their location on a knowledge or skill continuum. Such information can be obtained by examining the item's Rasch difficulty index.

This index estimates the difficulty level of an item with respect to the knowledge required to correctly respond to it (Wright & Stone, 1979). Extreme positive values describe hard items which challenge even the most able students. Extreme negative values describe easy items which even the least able students could probably answer correctly. A unique aspect of the index is its sample-free characteristic, that is, the estimation of item difficulty is independent of the specific abilities of the calibration sample.

Given this characteristic of the Rasch difficulty index, it is theoretically possible to identify a group of items that would divide the range of difficulties into equal intervals. This set of items would define an ordered, equal interval continuum, with each individual item identifying a unique location.

When working with most data sets, items often tend to cluster around a single level of difficulty. This results in a duplication of measurement at that level, while another level may be left unmeasured. Although unique, equally spaced item difficulties are uncommon, it is often possible to identify a set of items that approximate this ideal. The closer the approximation, the clearer the definition of the entire continuum.

A Rasch analysis is based on certain assumptions concerning what happens when a student responds to the items on a test. Before one can select items for inclusion on the final form of a test, a complete

analysis must evaluate how well these assumptions have been satisfied by the data (Wright & Stone, 1979). If the data satisfy the assumptions, they are said to "fit" the model; otherwise they do not.

Two assumptions must be evaluated: person-fit and item fit. The first concerns the plausibility of an individual's pattern of responses. Persons correctly answering all of the hard items would not be expected to miss numerous easy ones. If they did, however, their response patterns would be different from those proposed by the model. The severity of this "misfit" would determine whether that person would be considered in the analysis. The second issue focuses on the plausibility of the general pattern of responses to each item across all persons. A smaller proportion of low ability students, as compared to those with high ability, would be expected to correctly answer a very difficult question. If this expectation is not met, the item is not functioning as the model proposed it would. Again, the severity of this "misfit" would determine whether the item would be included in the analysis.

The items selected for the LTT were those that 1) fit the Rasch model properly, and 2) approximated an equally-spaced set of items on the ordered continuum of difficulty. The BICAL-Version 3 (Mead, Wright, & Bell, Note 3) computer program was used to obtain all of the latent-trait item analysis statistics.

Again, student responses to the 30 items selected for the LTT were culled from their responses on the UT. Scores for the LTT were calculated by summing each student's correct responses to this subset of items.

Sampled Test. The 30 items selected for the ST were chosen according to a stratified, random sampling procedure. Random sampling guaranteed that any differences between the sample and parent population were functions of chance. Stratifying guaranteed representation of the addition and subtraction subgroups.

The UT was stratified into addition and subtraction problems. A table of random numbers was then used to identify 15 items from each strata. These items comprised the ST. The scores for the ST were computed on the appropriate subset of items in a manner paralleling that used for the CT and LTT.

Test Administration

Teachers and principals expressed four major concerns regarding the testing of their students. First, several teachers advocated the separation of the addition and subtraction problems on the test. They felt the students' performances would be more indicative of their knowledge if this separation existed. Second, both teachers and principals indicated that several short testing periods would be less disruptive than a single, longer one. They also agreed the students would be more apt to concentrate under these conditions. Third, teachers and principals were concerned about the use of the test results for comparisons of classes or schools. It was agreed that no such comparisons would be made. Finally, the principals supported the anonymity of the students to the researcher. To do otherwise would have required an excessive amount of administrative time and effort.

These concerns were reflected in the administration of the UT. As mentioned earlier, the addition and subtraction items were separated onto four pages. The test was administered in four sections, one per page. One section was completed per day. Approximately 15-20 minutes were allocated for the completion of each section, and only student response data was collected.

The administration of the entire test spanned five consecutive days, with the fifth day reserved for make-up testing. About 80% of the sample took the test in the week beginning on November 15, 1982. The remaining students were tested during the week of December 6, 1982.

Each teacher received a packet containing a cover letter, test directions, and test booklets (see Appendixes A, B, and C for copies). The cover letter explained the purpose of the project, the test itself, and the testing dates. A set of explicit directions specified the administration procedures.

Data Analysis

The following discussion details the analyses used to examine the questions addressed by this study. Each question is discussed as follows: 1) a review of the specific question, 2) an operational definition of the purpose of the test, 3) an explanation of the specific analyses used, and 4) a statement of the statistical hypotheses tested.

Question 1. When the purpose of testing is the differentiation of students, how do the results of the tests compare with one another?

Consider several students, all having slightly different levels of knowledge of the subject area. It is often useful to identify the

better students, that is, to differentiate among students with respect to their relative levels of knowledge. To accomplish this, the variation in raw scores must be sufficient to separate all of the students. The most knowledgeable student would receive the highest score, while the least knowledgeable one would receive the lowest. All other scores would be spread between these extremes.

The variance of a group of scores estimates the amount of variation, or separation, present in the scores. Its value is large when the scores are heterogeneous, and small when they are homogeneous. Consequently, it provides an excellent estimate of the degree to which a test differentiates among students.

The focus of the analysis used to answer Question 1 compared the variances of the CT, LTT, and ST. A direct comparison of these variances with that of the UT was not considered due to the relationship between test length and variance. As test length increases, so does the potential for greater variance. The huge discrepancy between 30 and 110 item tests would have severely biased the outcome of this comparative analysis.

It would be interesting, however, to make such an "absolute" comparison of the four variances. To do so would require an estimation of the variance of the UT if it were scaled down to a 30 item test. The variance of the ST provides such an estimation. Due to the nature of the random sampling procedure used to select the items for the ST, any differences between the ST and UT are functions of chance. Therefore, a comparison of the variances of the CT and LTT with that of the

ST would estimate the extent to which these variances accurately reflect that of the UT.

A statistical test of equivalency of the variances of the CT, LTT, and ST empirically estimated the validity of each test for the purpose of differentiation. Three dependent t-tests were used to examine the following null hypotheses:

$$H_0: \sigma^2_{ct} = \sigma^2_{ltt},$$

$$H_0: \sigma^2_{ct} = \sigma^2_{st}, \text{ and}$$

$$H_0: \sigma^2_{ltt} = \sigma^2_{st}.$$

Question 2. When the purpose of testing is the location of students on a continuum of knowledge, how do the results of the tests compare with one another?

Consider now each student individually. It is useful to measure what each student knows, that is, to estimate his or her location on a continuum of achievement. To accomplish this, the continuum of achievement must be defined, and a person's test performance must be translated into a location on that continuum.

When test data are analyzed according to the tenets of the Rasch model, the achievement continuum is defined in terms of a set of δs (i.e., item difficulties). These δs can be combined with a student's raw score to describe their location on the continuum in terms of β (i.e., person ability).

A unique aspect of this analysis is that the estimates of δ are freed from the effects of β. Hence, an equivalent item difficulty is

hypothesized for any item common among tests. A Rasch analysis of the UT would provide an estimate of the δ associated with each of the 110 items. These δs can be used to estimate a β for each person taking the UT. Furthermore, providing that all items in the UT fit the model, any subset (i.e., CT, LTT, or ST) of δs from this calibration can be used to estimate βs for each person taking that subset of items.

It should be noted that it may not be possible to estimate the βs for certain students. Some of these students will have significant person-fit statistics; others will have zero or perfect scores on a given test. A detailed discussion of the reasons for these missing estimates is presented in Chapter 4.

The analysis used to answer Question 2 focused on a comparison of each individual's β as calculated on the UT, CT, LTT, and ST. A test of the statistical equivalency of the β_{ct}, β_{ltt}, and β_{st} to the β_{ut} empirically estimated the validity of each test for the purpose of locating students on a continuum of knowledge.

One of several appropriate statistical tests could have been chosen for the analysis of Question 2. However, a repeated measures ANOVA was intuitively more satisfactory than the correctional or regression analyses that were also considered. This analysis was used to examine the following hypothesis:

$$H_0: \mu_{\beta_{ut}} = \mu_{\beta_{ct}} = \mu_{\beta_{ltt}} = \mu_{\beta_{st}}.$$

Subsequent post-hoc comparisons between these means were to be performed when necessary.

Question 3. When the purpose of testing is the generalization of student performance to a universe of items, how do the results of

the tests compare with one another?

Again, consider each student individually. It is often useful to generalize student performance from a small set of items (i.e., a test) to an entire universe of items. To accomplish this, each student's performance must be estimated in such a way that it is independent of the number of items chosen from the domain.

The proportion of correct responses (i.e., p) estimates a student's test performance on a scale from 0 to 1. Its value is close to 1 when the student answers most of the items correctly, and close to 0 when most are answered incorrectly. It is easily calculated by dividing a student's raw score by the total number of items on the test.

The focus of the analysis used to answer Question 3 compared each individual's p as calculated on the UT, CT, LTT, and ST. A test of the statistical equivalency of the p_{ct}, p_{ltt}, and p_{st} to the p_{ut} empirically estimated the validity of each test for the purpose of generalizing student performance to a universe of items.

As with the analysis of the data pertaining to Question 2, several statistical tests could have been chosen for the analysis of Question 3. Again, a repeated measures ANOVA was used to examine the following hypothesis:

$$H_0: \mu_{p_{ut}} = \mu_{p_{ct}} = \mu_{p_{ltt}} = \mu_{p_{st}}.$$

Subsequent post-hoc analyses between these means were to be performed when necessary.

Statistical Significance

The results of any inferential test were considered significant

if they would have occurred by chance less than 5 times out of 100 (i.e., $\alpha = .05$). This significance level was adjusted according to Bonferroni's Inequality when repeated testing of statistics occurred.

Chapter Four

Results

Introduction

The objective of this study was to empirically examine the proposition that different item selection procedures contribute to the development of achievement tests that optimally served different purposes. To accomplish this, a test of 110 single-digit addition and subtraction items was administered to a sample of students. The results of this test provided the information needed to develop three additional 30 item tests. These tests consisted of items that were selected from the original 110 on the basis of their classical discrimination indexes, their Rasch item-fit and item difficulty statistics, or a random sampling design. Performances on all four tests were then compared with respect to the differentiation of students, location of students on a continuum of knowledge, or generalization of student performance to a universe of items. This chapter, which describes the results of these comparisons, is divided into six sections. The first provides a description of all four tests. The second, third, and fourth discuss the results of the inferential analyses used to examine Questions 1-3. The fifth summarizes the results of these analyses, and the final section examines the results of several secondary analyses.

Descriptive Information

The purpose of this section is to present a clear, unambiguous description of each of the four tests. It should be noted that the

item statistics and test results for each test were calculated in both a raw score and latent-trait metric. Verbal and empirical descriptions of the item statistics and test results are provided in both metrics.

UT. The UT consisted of 110, single-digit addition and subtraction problems. It defined a finite universe of items from which the other three tests were drawn. It was administered to a sample of 371 first grade students. Their responses to the items were scored as correct (1) or incorrect (0); hence the minimum possible score was 0 while the maximum was 110.

The statistics for each item on the UT are shown in Appendix D. Summary classical item difficulties, classical discrimination indexes, and Rasch item difficulties for the UT are found in Tables 3, 4, and 5 respectively.

As shown in the first row of Table 3, the classical difficulty indexes for the items on the UT ranged from .61 to .97, with a mean of .79. The standard deviation of the distribution of difficulties was .11, indicating the presence of a small amount of variation.

An examination of the first row of Table 4 indicates that the classical discrimination indexes ranged from .23 to .78, with a mean of .55. The standard deviation of this distribution was .13, indicating that the majority of the items adequately discriminated between high and low scoring students. Both the classical difficulty and discrimination indexes were calculated on the responses from all 371 students in the sample.

Finally, an examination of the first row of Table 5 indicates that the Rasch item difficulty indexes ranged from -3.28 to +1.77,

Table 3

Summary Item Statistics - Classical Difficulties

Test	N	Mean	S.D.	Range
UT	110	0.79	0.11	0.61 - 0.97
CT	30	0.72	0.03	0.67 - 0.78
LTT	30	0.84	0.11	0.61 - 0.97
ST	30	0.79	0.11	0.65 - 0.97

Note: All statistics have been rounded to two decimal places.

Table 4

Summary Item Statistics - Classical Discrimination Indexes

Test	N	Mean	S.D.	Range
UT	110	0.55	0.13	0.23 - 0.78
CT	30	0.71	0.03	0.67 - 0.78
LTT	30	0.49	0.13	0.23 - 0.71
ST	30	0.56	0.15	0.23 - 0.78

Note: All statistics have been rounded to two decimal places.

Table 5

Summary Item Statistics - Rasch Difficulties

Test	N	Mean	S.D.	Range
UT	110	0.00	1.35	-3.28 - +1.77
CT	30	1.06	0.29	+0.57 - +1.67
LTT	30	-0.75	1.52	-3.28 - +1.77
ST	30	-0.03	1.38	-2.82 - +1.47

Note: All statistics have been rounded to two decimal places.

with a mean of 0. The standard deviation of this distribution of difficulties was 1.35. It should be remembered that extreme negative values for this index describe easy items that even the least able students can probably answer correctly, while extreme positive values describe hard items that challenge even the most able students.

An examination of the Rasch item-fit statistics found in Appendix D indicated a range from -1.27 to +5.76 (see column 5). These fit statistics were individually calculated for each item on the UT. They are empirical estimations of the extent to which the observed proportions of correct answers for different ability groups differ from those proportions expected by the Rasch model. For example, an increase in the proportion of correct answers to a difficult item is expected when moving from less able to more able score groups. If, however, the less able score groups have a substantially higher proportion of correct answers than would be expected given their ability or achievement level, the item does not function as expected. The severity of this "misfit" of observed and expected proportions would determine whether the item should be deleted from the test. All of the fit statistics for the 110 items on the UT were judged to be within acceptable limits.

Both the Rasch item difficulty indexes and item-fit statistics were calculated on the responses of 343 students, reflecting the deletion of 28 "non-measured examinees." These latter students were identified by discrepant person-fit statistics, scores of 0 or scores of 110. These characteristics are fully described in the following discussion of Rasch ability estimates. It should be noted, however,

that the responses of all of the students were included in the calculation of any statistics in the raw score metric.

The raw scores and Rasch ability estimates for each student are shown in Appendix E. Summary statistics for the raw scores and Rasch ability estimates for the UT, CT, LTT and ST are found in Tables 6 and 7 respectively. It should be noted that the ABILEST computer program (Ryan, Note 5) was used to generate all Rasch ability estimates.

An examination of Appendix E indicates that Rasch ability estimates were not computed for numerous students. Some of these missing estimates were associated with students who had excessive person-fit statistics. Person-fit statistics are similar in nature to the item-fit statistics discussed earlier. They are empirical estimations of the extent to which observed patterns of item responses differ from the patterns expected by the Rasch model. For example, a student is expected to answer easy items correctly and hard ones incorrectly. If, however, a student's response pattern indicated that the hard items were answered correctly and the easy ones answered incorrectly, the observed response pattern would not coincide with what was expected. If this discrepancy between the observed and expected patterns is substantial, the validity of the student's performance cannot be established, and measurement of the student's ability cannot be made.

Rasch ability estimates were not computed for several other students because they received zero or perfect test scores. These scores reflect the lower and upper bounds of the ability continuum measured by a test. In the case of a student who received a score of zero, it is impossible to determine the extent to which their true ability is below the lowest level measured by the test. For the

Table 6

Summary Test Statistics - Raw Scores

Test	N	Mean	Variance (S.D.)	Range	KR20	SEM
UT	371	86.46	588.55 (24.26)	3-110	.98	3.43
CT	371	21.01	111.30 (10.55)	0-30	.98	1.49
LTT	371	25.21	30.58 (5.53)	1-30	.91	1.64
ST	371	23.67	48.16 (6.94)	1-30	.93	1.84

Table 7

Summary Test Statistics - Rasch Abilities

	Uncorrected Sample				Corrected Sample			
Test	N	Mean	S.D.	Range	N	Mean	S.D.	Range
UT	332	2.29	1.75	-4.45 to +5.26	199	1.75	1.41	-2.89 to +4.55
CT	258	2.22	1.96	-2.26 to +4.31	199	1.76	1.98	-2.26 to +4.31
LTT	264	1.70	1.51	-4.82 to +3.38	199	1.68	1.39	-4.07 to +3.38
ST	284	1.90	1.65	-4.19 to +3.79	199	1.73	1.48	-3.40 to +3.79

Note: The uncorrected sample consisted of all of the students with ability estimates for each individual test, while the corrected sample consisted of only those students with estimates for all four tests.

student with a perfect score, it is impossible to determine the discrepancy between their true ability and the highest ability measured by the test. Consequently, the abilities of the students receiving these scores cannot be accurately established.

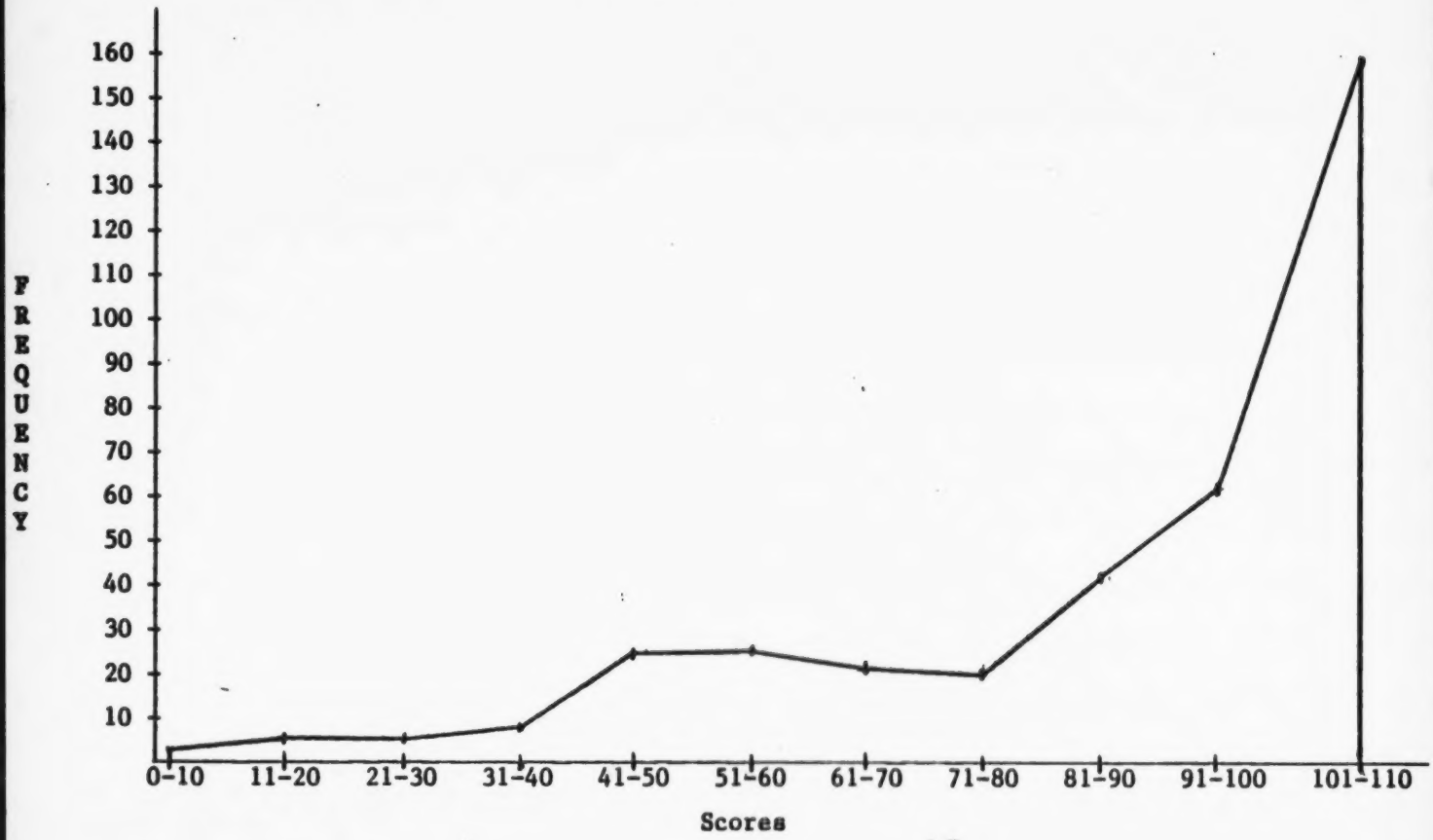
An examination of Table 6 indicates that the UT was a sufficiently reliable, fairly easy test containing a substantial amount of variation among students. The KR20 for this administration of the test was .98, and the SEM was 3.43. The average student was able to correctly answer about 79% of the items. As shown in Figure 1, the distribution of scores was negatively skewed, with approximately 45% of the students scoring above 100.

An examination of Table 7 indicates that most students were measured by the UT. Only 27 students with misfitting response patterns and 12 with scores of 110 were eliminated.[1] The UT measured a wide range of student ability, with the average student demonstrating adequate knowledge of more than two-thirds of the entire difficulty continuum. The abilities of the students measured with the UT ranged from -4.45 to +5.26, with a mean of 2.29 and standard deviation of 1.75.

It should be noted that these "uncorrected" summary statistics should not be compared across tests due to the varying nature of the samples of students measured by each test. For example, in the case of the CT, about 6% of the lowest and 17% of the highest ability students were not considered in the computation of the summary statistics. When contrasted to the UT sample in which only 3% of the highest ability students were eliminated, the differences between

Figure 1

Distribution of Raw Scores on the UT
(n=371)



the samples on which the computations were based is quite evident.

Consequently, a second set of summary statistics was computed on the Rasch ability estimates (see Table 7, columns headed by "corrected sample"). These statistics were based on those students for which ability estimates were available for all four tests. Although this "corrected" sample only consisted of 199 students, it provided a theoretically sound basis upon which the ability estimates for all four tests would be compared. An examination of these statistics indicates the mean ability for those students measured by all four of the tests was 1.75, with a standard deviation of 1.41.

Item Selection. The results of the UT provided the information needed to develop the CT, LTT, and ST. The CT consisted of the 30 items from the UT having the highest classical discrimination indexes. The LTT included the 30 items that best reflected the entire continuum of Rasch item difficulty defined by the items on the UT. The ST consisted of 30 items randomly selected from those on the UT.

The sets of items selected for each test are identified in Table 8. An examination of the items on these tests indicated that the CT consisted of subtraction items only. Only subtraction items adequately differentiated between high and low achieving students. The LTT included 19 addition and 11 subtraction items, and the ST included 15 items of each type. Furthermore, several items were included on more than one test. The CT and LTT had 5 items in common, the CT and ST had 11, and the LTT and ST had 8. Of these duplications, 19 involved subtraction items, and 5 involved addition items.

Table 8

Items Included on the CT, LTT, and ST

CT	LTT	ST
S059	A001	A004
S060	A003	A008
S062	A016	A010
S065	A017	A011
S066	A019	A017
S071	A020	A022
S072	A021	A023
S073	A022	A029
S079	A025	A034
S080	A026	A038
S081	A028	A041
S083	A030	A042
S084	A033	A044
S085	A037	A050
S087	A038	A055
S089	A040	S056
S092	A050	S065
S093	A053	S070
S094	A055	S075
S095	S059	S079
S097	S060	S081
S098	S065	S083
S100	S066	S086
S101	S070	S087
S102	S072	S092
S104	S075	S093
S105	S076	S097
S106	S077	S098
S107	S107	S104
S108	S110	S107

Note: The "A" prefix refers to addition items, and the "S" prefix refers to subtraction items.

CT. The relevant data pertained to the CT are included in Table 3, p. 36. The classical difficulty indexes for the CT tended to be lower and more restricted in range than those for the UT, LTT, or ST. These indexes ranged from .68 to .78, with a mean of .72. The standard deviation of the distribution of difficulties was .03.

An examination of Table 4 (p. 36) indicates that the classical discrimination indexes for the items on the CT tended to be substantially higher and more restricted in range than those for any other test. These results are not surprising given the nature of the classical selection procedure. All of these indexes were larger than 0.66, indicating that all of the selected items were excellent choices relative to the selection criteria established for the CT in Chapter 3. The mean discrimination index for the CT was .71, and the standard deviation associated with the distribution of indexes was .03.

An examination of Table 5 (p. 36) indicates that the Rasch difficulty indexes also tended to be substantially higher and more restricted in range than those for the UT, LTT, or ST. These difficulties ranged from .57 to 1.67, with a mean difficulty of 1.06. The standard deviation of the distribution of difficulties was .29.

The summary statistics associated with the raw scores on the CT can be found in Table 6 (p. 39). They indicate that the CT was a reliable, relatively difficult test containing a substantial amount of variation among students. The KR20 was .98, and the SEM was 1.49. This was the smallest SEM for any test. The average student was able to correctly answer about 70% of the questions. This number represents the smallest proportion of correct answers for any of the tests.

As expected, the most unique characteristic of the CT was its ability to differentiate among students. The variance for the distribution of raw scores was 111.30, more than twice that of the ST and three-times that of the LTT.

The summary statistics for the Rasch ability estimates associated with the CT can be found in Table 7 (p. 40). They indicate that fewer students were measured with this test than with any other. Of the unmeasured students, 27 had misfitting response patterns, 22 had scores of 0, and 64 had perfect scores of 30. It is interesting that none of the students received a score of 0 on any other test. An examination of the "corrected" statistics in Table 7 indicates that the students' abilities as measured by the CT tended to be less restricted but very comparable in level to those for the UT, LTT, or ST.

To summarize, the CT appeared to function as expected. It was the most difficult subtest, with the items being harder and more discriminating than those on the LTT or ST. Raw score performances on the test were slightly lower than those for other tests, yet the abilities of the students were quite comparable to those found on the other tests.

LTT. The relevant data pertaining to the LTT also were included in Table 3, p. 36. The classical item difficulty indexes for the LTT tended to be higher than those for any other test, that is, the items were somewhat easier. However, these indexes appeared to be distributed in a similar fashion to those of the UT or ST. The indexes ranged from .61 to .97, with a mean of .84. The standard deviation of the distribution of difficulties was .11.

An examination of Table 4 (p. 36) indicates that the classical discrimination indexes for these items tended to be similar to those found for the UT or ST. The mean discrimination index was .49, with a standard deviation of .13. Eight of these indexes were below the classical selection criterion of .40.

An examination of Table 5 (p. 36) indicates that the Rasch item difficulty indexes tended to be slightly lower and more varied than those for any other tests, again indicating slightly easier items. Given the nature of the latent-trait selection procedures, these results were expected. The mean difficulty was -.75 with a standard deviation of 1.52.

A more detailed examination of these Rasch difficulties indicated that the items were excellent choices relative to the selection criteria established for the LTT. As previously indicated, the first criterion concerned the proper fit of the item to the Rasch model. It should be remembered that an item-fit statistic empirically estimates this fit issue, that is, the discrepancy between the observed and expected proportions of correct answers across different ability groups. An examination of Appendix D indicates that the fit statistics for the items on the LTT ranged from -.80 to +3.82. Consequently, all of the items were judged to fit the Rasch model.

The second criterion concerned the extent to which the LTT items approximated an ordered, equally-spaced difficulty continuum. A hypothetical continuum was defined by dividing the range of Rasch item difficulties on the UT into 30 equal segments. The easiest UT item established the lower limit of this continuum, while the hardest

Table 9

Descriptions of Rasch Item Difficulty Continuums for the CT, LTT, and ST

Theoretical Difficulty	CT Observed Difficulty	CT Discrepancy	LTT Observed Difficulty	LTT Discrepancy	ST Observed Difficulty	ST Discrepancy
1.77	1.67	-.10	1.77	-	1.47	-.30
1.60	1.67	.07	1.59	-.01	1.41	-.19
1.42	1.47	.5	1.41	-.01	1.39	-.03
1.25	1.39	.14	1.22	-.03	1.35	.10
1.07	1.35	.32	1.07	-	1.22	.15
.90	1.28	.38	.91	.01	1.22	.32
.73	1.28	.55	.77	.04	1.11	.38
.55	1.22	.67	.57	.02	1.11	.56
.38	1.22	.84	.33	-.05	1.07	.69
.20	1.18	.98	.19	-.01	1.04	.84
.03	1.11	1.08	.02	-.01	1.04	1.01
- .15	1.11	1.26	-.17	.02	.91	1.06
- .32	1.11	1.43	-.33	.01	.81	1.13
- .49	1.07	1.56	-.58	.09	.59	1.08
- .67	1.07	1.74	-.65	-.02	.52	1.19
- .84	1.04	1.88	-.81	-.03	.19	1.03
-1.02	1.04	2.06	-1.03	.01	.02	1.04
-1.19	1.04	2.23	-1.17	-.02	-.01	1.18
-1.36	1.00	2.36	-1.32	-.04	-.04	1.32
-1.54	.93	2.47	-1.55	.01	-.26	1.28
-1.71	.91	2.61	-1.80	.09	-.58	1.13
-1.89	.86	2.75	-1.87	-.02	-.81	1.08
-2.06	.81	2.87	-1.95	-.11	-.98	1.08
-2.24	.81	3.05	-2.28	.04	-1.08	1.16
-2.41	.79	3.20	-2.37	-.04	-1.27	1.14
-2.58	.79	3.37	-2.58	-	-1.95	.63
-2.76	.77	3.53	-2.82	.06	-2.28	.48
-2.93	.62	3.55	-2.82	-.11	-2.69	.24
-3.11	.59	3.70	-2.96	-.15	-2.69	.42
-3.28	.57	3.85	-3.28	-	-2.82	.46
		Σ = 54.91		Σ = -.26		Σ = 21.66

Note: These theoretical difficulties were determined by dividing the range of Rasch item difficulties on the UT into 30 equal segments. The observed difficulties are those calculated on the UT for the appropriate items on the CT, LTT, and ST. The discrepancies represent the difference between the observed and theoretical difficulties.

item defined the upper limit. The additional 28 points were equally spaced between these limits. The difficulty continuums defined by the items on the CT, LTT, and ST were then compared to this "theoretical" continuum. As shown in Table 9, not only did the LTT items closely approximate the entire "theoretical" continuum, but they did so in a much more efficient manner than the CT or ST. Again, given the nature of the latent-trait selection procedure, these results were not surprising.

The summary statistics associated with the raw scores on the LTT can be found in Table 6 (p. 39). They indicate that the LTT was a reliable, relatively easy test containing a relatively small amount of variation among students. The KR20 was .91, and the SEM was 1.64. The average student correctly answered about 84% of the items. This number represents the largest proportion of correct answers for any test. The variance for the distribution of raw scores on the LTT was 30.58. This number represents the smallest variation among students for any test.

The summary statistics for the Rasch ability estimates associated with the LTT can be found in Table 7 (p. 40). They indicate that the abilities of 264 students were measured with the LTT. This reflected the elimination of 27 students with misfitting response patterns and 80 students with perfect scores of 30. No students received a score of 0 on the LTT. The "corrected" mean ability was 1.68, with a standard deviation of 1.39. These were quite consistent with those computed for the other tests.

To summarize, the LTT appeared to be an easier test than the CT or ST. The items spanned the entire range of item difficulty; consequently, both easy and hard items were included on the test. Student performance on the LTT was slightly higher than on the other tests, yet measured abilities were consistent with those found on the UT, CT, or ST.

ST. The relevant data pertaining to the ST are displayed in Table 3, p. 36. The classical item difficulty indexes for the ST tended to fall between those of the CT and LTT. They were a direct reflection of the difficulties found on the UT, with a mean of .79 and standard deviation of .11.

An examination of Table 4 (p. 36) indicates that the classical discrimination indexes for these items also tended to fall between those of the CT and LTT. They also tended to reflect those found on the UT, with a mean of .56 and standard deviation of .15. The range of the indexes was .23 to .78, slightly greater than that of the LTT and substantially greater than that of the CT.

An examination of Table 5 (p. 36) indicates that the Rasch item difficulties tended to fall between those of the CT and LTT. They closely approximated those of the UT, with a mean of -.03 and standard deviation of 1.38. The range of difficulties was -2.82 to +1.47, slightly more restrictive than that of the LTT but much more varied than that of the CT.

The summary statistics associated with the raw scores on the ST can be found in Table 6 (p. 39). They indicate the ST was a reliable, easy test containing a moderate amount of variation among students.

The KR20 was .93, and the SEM was 1.84. This was a higher SEM than that for the CT or LTT. The average student was able to correctly answer about 79% of the items on this test. This approximated the proportion of correct answers for the UT, but was greater than that for the CT and less than that for the LTT. The variance for the distribution of raw scores on the ST was 48.16. This was less than one-half of the variance of the CT, but slightly more than that associated with the LTT.

The summary statistics for the Rasch ability estimates associated with the ST can be found in Table 7 (p. 40). They indicate that more students were measured by this test than by the CT or LTT. Of the unmeasured students, 27 had misfitting response patterns and 60 had perfect scores of 30. As with the LTT, no students received a score of 0 on the ST. The corrected mean ability was 1.73 with a standard deviation of 1.48. These statistics were very similar to those computed for the other tests.

As was expected, the ST was quite representative of the UT. The ST and UT tended to be exact duplications of one another, regardless of the statistics on which they were compared. Consequently, it appeared as though there was a substantial amount of empirical support for the use of a stratified random selection procedure.

To summarize, the items on the ST tended to be less difficult than those on the CT, but more difficult than those on the LTT. Raw score performances were balanced between those on the CT and LTT, and student abilities as measured by the test were comparable to those measured by the UT, CT, and LTT.

Analysis of Question 1

Question 1: When the purpose of testing is the differentiation of students, how do the results of the tests compare with one another?

The analysis for Question 1 focused on the comparison of the variances of the CT, LTT, and ST. Three variances were computed from the students' raw scores on the CT, LTT, and ST.

Three dependent t-tests were used to determine whether or not the observed differences among variances were statistically significant. The following null hypotheses were tested:

$$H_0: \sigma_{ct}^2 = \sigma_{ltt}^2,$$

$$H_0: \sigma_{ct}^2 = \sigma_{st}^2, \text{ and}$$

$$H_0: \sigma_{ltt}^2 = \sigma_{st}^2.$$

The t-tests were performed on the data from all 371 students in the sample. The results were as follows:

$$t_{\sigma_{ct}^2 \sigma_{ltt}^2} = 23.41,$$

$$t_{\sigma_{ct}^2 \sigma_{st}^2} = 21.11, \text{ and}$$

$$t_{\sigma_{ltt}^2 \sigma_{st}^2} = -10.45.$$

Significance levels were adjusted from .05 to .017, reflecting the application of Bonferroni's Inequality for three repeated testings. All three t-statistics were significant at this level. Consequently, all three null hypotheses were rejected.

As mentioned in Chapter 3, the variance of the ST was assumed to be representative of that of the UT. An examination of Table 10 indi-

Table 10

Raw Score Correlations Between Tests

		TEST			
		UT	CT	LTT	ST
T	UT	1.00	.92	.95	.98
E	CT		1.00	.80	.92
S	LTT			1.00	.91
T	ST				1.00

cates that the correlation between raw scores on the UT and ST is .98. This correlation implies that 96% of the variance in the UT can be accounted for by the variance in the ST. Furthermore, the ST and UT were quite similar relative to the descriptive statistics associated with classical item difficulties, classical discrimination indexes, Rasch item difficulties, proportion of correct answers, and Rasch ability estimates. It appeared to be safe to assume that the ST variance estimate was an unbiased estimate of the UT variance if the UT was composed of 30 items.

In summary, the CT yielded greater differences among students than the LTT or ST. Furthermore, the LTT yielded smaller differences among students than did the ST. It can be concluded that tests developed according to classical selection procedures tend to overestimate the differences among students, while those developed by latent-trait selection procedures underestimate such differences.

Analysis of Question 2

Question 2: When the purpose of testing is the location of students on a continuum of knowledge, how do the results compare with one another?

The analysis for Question 2 focused on the comparison of student's Rasch ability estimates (i.e., βs) as estimated by their performances on the UT, CT, LTT, and ST. The analysis required the calibration of Rasch difficulties for each item on the UT.[2] The appropriate sets of difficulties were then used to estimate the Rasch abilities associated with the raw scores on the UT, CT, LTT, and ST.

A repeated measures ANOVA was used to determine whether any of the observed differences in mean abilities were statistically significant. The following null hypothesis was tested:

$$H_0: \mu_{\beta_{ut}} = \mu_{\beta_{ct}} = \mu_{\beta_{ltt}} = \mu_{\beta_{st}}.$$

The analysis was performed on the data from the 199 students having ability estimates for all four tests.[3] The results can be found in Table 11. They indicate a lack of any statistically significant differences ($\alpha = .05$) between the mean Rasch abilities on the UT, CTT, LTT, and ST. Therefore, the null hypothesis was accepted.

In summary, neither the UT, CT, LTT, or ST provided statistically different ability estimates for the sample of students. It can be concluded that tests developed according to the classical, latent-trait, or sampling procedures examined in this study are equally effective at locating students on a continuum of knowledge when estimates of student ability along the continuum are based on the Rasch model.

Analysis of Question 3

Question 3: When the purpose of testing is the generalization of student performance to an item universe, how do the results of the test compare with one another?

The analysis for Question 3 focused on the comparison of the proportions of correct answers each student received on the UT, CT, LTT, and ST. These proportions were calculated by dividing the sum of a student's correct answers by the total number of items on the test.

A repeated measures ANOVA was used to determine whether any of the observed differences in mean proportion were statistically signifi-

cant. The following null hypothesis was tested:

$$H_0: \mu_{P_{ut}} = \mu_{P_{ct}} = \mu_{P_{ltt}} = \mu_{P_{st}}.$$

The analysis was performed on the data from all 371 students in the sample. The results can be found in Table 12. They indicate the existence of at least one statistically significant difference among proportions. Therefore, the null hypothesis was rejected.

A post-hoc analysis of tne data examined all pair-wise comparisons of the four means using Tukey's procedures. The computed Q-statistics were as follows:

$$Q_{ut-ct} = 15.97,$$

$$Q_{ut-ltt} = -10.18,$$

$$Q_{ut-st} = -.57,$$

$$Q_{ut-ltt} = -26.15,$$

$$Q_{ct-st} = -16.54, \text{ and}$$

$$Q_{ltt-st} = 9.61.$$

The critical value for Q at the .05 significance level with 4 and 1480 df was 3.63. All of the computed values of Q were statistically significant with the exception of that associated with the comparison of the mean proportions between the UT and ST.

In summary, the CT yielded smaller proportions of correct answers than did the UT, LTT, or ST. Furthermore, the LTT yielded larger proportions of correct answers than the UT, CT, or ST. The proportions of correct answers for the UT and ST were statistically equivalent. It can be concluded that the tests developed according to the classical

Table 11

Repeated Measures ANOVA for Rasch Abilities

Source	SS	df	MS	F	PROB.
Between Test	.67	3	.22	.57	.64
Test * Subject Interaction	232.41	594	.39		

Table 12

Repeated Measures ANOVA for Proportions of Correct Responses

Source	SS	df	MS	F	PROB.
Between Test	3.74	3	1.25	117.37	<.001
Test * Subject Interaction	11.79	1110	.01		

procedures examined in the study tend to underestimate the true proportion of correct answers, while tests developed according to the latent-trait procedures tend to overestimate the true proportions. Furthermore, tests developed by random selection procedures tend to accurately reflect the true proportions of correct answers.

Summary of the Results of the Primary Analyses

The results of all the analyses are summarized in Table 13. Several conclusions can be drawn from an examination of this table. Relative to test purpose, it is apparent that significant differences do not exist between the item selection procedures if the purpose of testing is the location of students on a continuum of knowledge and if the estimates of location on the continuum are based on Rasch ability estimates. Contrary to this finding, significant differences between selection procedures do exist when the purpose of testing is either the differentation of students or the generalization of test results.

From a test development perspective, it is apparent that the ST provides an unbiased estimate of the UT, regardless of test purpose. In contrast, both the CT and LTT provide biased estimates of the variance and proportion of correct answers found on the UT. In the case of the CT, this bias takes the form of an overestimation of the variance and an underestimation of the proportion of correct answers. For the LTT, the bias is evidenced by an underestimation of the variance and an overestimation of the proportion of correct answers.

Secondary Analysis of the Data

An examination of the results of the primary analyses indicated

Table 13

A Summary of the Comparisons of the UT with the CT, LTT, and ST

		UT		
		σ^2	μ_β	μ_p
TEST	CT	Overestimation	-	Underestimation
	LTT	Underestimation	-	Overestimation
	ST	-	-	-

Note: The entries in this table reflect the comparison of the σ^2, μ_β, and μ_p of the UT with those of the CT, LTT, and ST. A "-" indicates the lack of a statistically significant difference. "Overestimation" indicates a significantly larger statistic on the CT, LT, or ST, and "underestimation" indicates a significantly smaller one.

that the CT and LTT provided biased estimates of the UT for two of the three test purposes. In an attempt to understand the nature of these biases, attention was focused on those students in the lower portion of the distribution. It is possible, for example, to both increase variance and lower the mean of a distribution of scores by decreasing the scores of the lower level students. Similarly, a decrease in variance and an increase in the mean could be obtained by increasing the scores of these students. A secondary analysis of the data was performed in an attempt to verify the proposition that the bias found in the CT and LTT could be attributed to the performance of the lower level students.

Students were divided into three groups on the basis of their relative standings on the UT. The lowest and highest 25% of the students were placed in Groups 1 and 3 respectively. The middle 50% were placed in Group 2. Group 1 consisted of 93 students who scored at or below 71, Group 2 of 185 students who scored between 72 and 104, and Group 3 of 93 students who scored at or above 105. It should be mentioned that it is common educational policy and practice to divide a distribution of students into lower, middle, and upper quartiles.

Since all of the biases found in the CT and LTT were observed in a raw-score metric (i.e., variance or proportion of correct answers), this metric was used for the secondary analysis of the data. The mean proportions of correct answers were calculated on each test for each of the three groups.

The descriptive statistics summarizing each group's performance on the UT, CT, LTT, and ST can be found in Appendix F, Table 1. An

examination of these data appeared to verify the proposition examined by this analysis. For Group 1, the mean proportions of correct answers for the UT and ST were .45 and .44 respectively. The mean proportions for the CT and LTT were .16 and .58 respectively. Similar, but less extreme differences appeared to exist for Group 2, and very small differences were observed for Group 3.

A repeated measures ANOVA was used to compare the means for each group across all tests. The results of these analyses indicated significant differences between the test means for all groups, and Tukey's post-hoc procedures were used to examine all pair-wise comparisons. The results of these analyses can be found in Appendix F, Tables 2 through 4.

The results of the secondary analysis are summarized in Table 14. These results tended to support the proposition that the bias found in the CT and LTT can be attributed to the performance of the lower quartiles of students. The CT severely underestimated the performance of Group 1, and slightly underestimated that of Group 2. Such underestimations would result in an increase in the variance and a decrease in the mean of a distribution of scores. The LTT severely overestimated the performance of Group 1, slightly overestimated that of Group 2, and very slightly overestimated that of Group 3. Although this indicates that some of the LTT bias can be attributed to Group 3, most of the bias originates with Group 1.

Table 14

A Summary of the Results of the Secondary Analysis

Comparison	Group		
	1	2	3
UT-CT	CT Underestimates	CT Underestimates	-
UT-LTT	LTT Overestimates	LTT Overestimates	LTT Overestimates
UT-ST	-	-	-

Chapter Five

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

Introduction

This chapter contains a brief summary of the study, a summary of the findings, a discussion of some considerations that may have affected the outcomes of the research, a discussion of the distribution of the test scores, a presentation of the conclusions drawn from the research, a list of recommendations for future research, and a discussion of the implications for educational practice. The chapter is divided into seven sections, one section for each of these topics.

Summary of the Study

Achievement tests have been used to differentiate among students, locate students on a continuum of knowledge, or generalize student performance from a small set of items to a universe of items. It seemed reasonable that the use of three different item selection procedures would contribute to the development of achievement tests that would best serve these purposes. Although substantial theoretical support for this proposition was evident, little empirical support was found. Consequently, the purpose of this study was to empirically examine the extent to which three item selection procedures contributed to the development of tests that optimally served the three purposes mentioned above.

To accomplish this, three sets of items were selected from a universe of 110 items measuring students' knowledge of single-digit addition and subtraction. The first set included those items with the

highest classical discrimination indexes. The second consisted of those items that best described the entire continuum of knowledge based on their Rasch item difficulty estimates. Finally, the third set included 30 randomly selected items. Considering the performances of 371 students on the three sets of 30 items as well as the set of 110 items, the study examined which set of items provided the most accurate results if the purpose of testing was the differentiation of students, location of students on a continuum of knowledge, or generalization of student performance from a small set of items to a larger universe of items.

Summary of Findings

The results of the analyses of the data revealed that significant differences did not exist between tests developed according to the three item selection procedures when the purpose of such testing was the location of students on a continuum of knowledge. The Rasch ability estimates for all of the students on all of the tests were quite comparable to one another. However, significant differences between the tests developed by the three selection procedures did exist when the purpose of testing was either the differentiation of students or the generalization of test results.

When considering the differentiation of students, the test developed according to the classical criteria overestimated the true variance of the distribution of test scores. This was apparently caused by a severe underestimation of the performance of the students in the lower quartile, and a slight underestimation of the performance of the students in the middle quartiles. In contrast, the test developed according to the latent-trait criteria underestimated the true variance.

This was apparently caused by a severe overestimation of the performance of the students in the lower quartile, a slight overestimation in the middle quartiles, and a very slight overestimation in the upper quartile.

When considering the generalization of test results, the test developed according to classical criteria underestimated the true proportion of correct answers of the students. The test developed according to latent-trait criteria overestimated the proportion. These results can also be attributed to the performance of the lowest quartile of students.

A Potential Reason for the Non-Significant Results Pertaining to the Estimation of Student Ability

The results of this study indicated non-significant differences among the tests when measuring student abilities, but significant differences when differentiating students or generalizing test results. A secondary analysis of the data provided an explanation as to the cause of the observed significant differences. At this point, it was reasonable to speculate as to the cause of the observed non-significant differences.

The non-significant difference among tests was found when the purpose of such testing was the location of students on a continuum of knowledge. In order to accomplish this purpose, it was necessary to 1) define the continuum of knowledge, and 2) locate students on it.

In this study, the continuum was defined by the difficulty levels of the items on the UT. Easy items, which even the least able

students could answer successfully, defined the lower end of the continuum. Hard items, which challenged even the most able students, defined the upper end. In order to locate a student on this continuum, it was necessary to obtain their responses to some of the items defining the continuum. If it was determined that the student's responses added up to a position on the continuum, the most reasonable place to estimate their location was where their responses changed from mostly correct to mostly incorrect. This location was considered the best estimate of a given student's ability.

Such an ability estimate takes into consideration the position of the items along the difficulty continuum. Hence, it was not unusual for each individual to have obtained statistically equivalent ability estimates, regardless of the test. The non-significant results found in the comparison of student abilities across tests could be directly attributed to the fact that such ability estimates transcend the tests upon which they are based.

It should be noted that the items used in this study were targeted toward the students responding to them, that is, the difficulty levels of the items were appropriately matched to the ability levels of the students. As long as this condition exists, similar results to those found in this study are expected.

Discussion of Distributional Issues

The distribution of scores on the UT was quite negatively skewed. Although such a distribution is not unusual for mastery or competency tests, it does warrant consideration in light of the results of the study.

A negatively skewed distribution indicates that the majority of the students found the test quite easy. In fact, 25% of the students correctly answered 105 or more of 110 items on the UT. As a consequence, many of the confounding factors (i.e., guessing, noise, etc.) affecting the test's reliability were minimized. This resulted in an extremely accurate assessment of student performance. Obviously, this accuracy contributed to a clean analysis and interpretation of the data.

It should be noted, however, that a similar situation may not have existed if the distribution of scores on the UT was normal or positively skewed. The results of this study should not be generalized to such distributions without further research examining the issue.

Conclusions

In considering the results of this study, it appears that different item selection procedures are differentially valid for two of the three different purposes. First, if the purpose of testing is the differentiation of students, the selection of items on the basis of their classical discrimination indexes will tend to differentiate the better and poorer students more effectively than would a test selecting the items according to latent-trait or stratified, random sampling procedures. Second, if the purpose of testing is the generalization of test results, the random selection of items will tend to more accurately reflect a student's proportion of correct answers to the universe of items than would a test selecting the items by classical or latent-trait procedures. It should be noted that if the purpose of testing is the location of students on a continuum of knowledge, each of the

three selection methods will tend to accurately accomplish this purpose providing the difficulties of the items are estimated using a one-parameter latent-trait model.

Three conclusions regarding the use of differentially developed tests can also be drawn from the results of this study. First, the selection of a set of items with high classical discrimination indexes will tend to underestimate the scores of students in the lower quartile on the entire set of items. The consequences of this underestimation are a decreased mean test score and an increased variance in the distribution of scores. Second, the selection of items that accurately reflect an entire ordered continuum of difficulty will tend to overestimate the scores of low level students on the entire set of items. The consequences of this overestimation are an increased mean test score and a decreased variance in the distribution of scores. Finally, a set of randomly selected items will tend to accurately reflect the scores of all students on the entire set of items, regardless of the quartile in which they are placed.

A final conclusion concerns the manipulation of the performance of lower quartile students. The results of the study indicated that if a set of selected items produced a biased estimate of a statistic, the bias would always be attributed to the scores of those students in the lower quartile.

Recommendations for Future Research

The results of this study provide empirical evidence in support of the proposition that tests developed according to different selection procedures optimally serve different purposes. However, it should be noted that these results were highly dependent on the information pro-

vided by the UT. An examination of the distribution of scores on this test indicated it was significantly negatively skewed. The shape of this distribution may have had a greater influence on the results of the analysis of Question 2 than did the effectiveness of either item selection procedure. Consequently, the results of this study may not be generalizable to sets of items with normal or positively skewed distributions.

It should also be noted that only three, very specific item selection procedures were examined. The criteria established for the classical, latent-trait and sampling procedures was intended to represent the criteria normally used to develop norm-referenced, criterion-referenced, and domain-referenced tests. However, neither the procedures nor the criteria were by any means intended to be exhaustive.

As a consequence, future research focusing on these distributional and selection procedure concerns should be conducted. Such work would contribute to a fuller, clearer understanding of the relationship between item selection procedures and test purposes.

Implications for Educational Practice

From a practical standpoint, the study has two implications. First, the relationship between three item selection procedures and test purposes has been clarified. Such clarification can have a direct effect on the development and use of tests. If, for example, it was necessary to develop a test that was to be used to separate the better from the poorer students, a set of items with high classical discrimination indexes would be selected. This test would effectively and efficiently accomplish the intended purpose. Consider, however, the

same test being used to evaluate students in terms of their proportions of correct answers. The results of this study indicate that the true performances of the lower level students would be severely underestimated. It would be highly unlikely for these students to appreciate the use of such a test. Consider a final example in which the same test was used to measure a student's location on a defined continuum of knowledge. If the level of difficulty for any of the items was appropriate for a student, the test would accurately reflect their ability, regardless of their relative standing in the group.

A second important practical implication concerns the performance of those students in the lower quartile. Many of the federal policies on the evaluation and funding of educational programs are inextricably connected with the performance of these students. This study has demonstrated that their performances can be severely biased by the selection of specific types of items. For example, if items were selected according to classical criteria, a student in the lower quartile would have to demonstrate a large increase in their knowledge of the domain of items before any progress could be measured by this specific set of items. In this particular study, a student in the lower quartile could conceivably have improved their UT score from 0 to 65, yet their CT score would not reflect any progress. Basing an evaluation of a program on the results of such tests could be quite misleading.

Reference Notes

1. Ryan, J.P. Personal communication, April, 1982.

2. Huynh, H., & Saunders, J.C. Solutions for some technical problems in domain referenced testing. Columbia, S.C.: University of South Carolina, 1980.

3. Mead, R.J., Wright, B.D., & Bell, S. R. BICAL-Version 3. Chicago: University of Chicago, 1979.

4. Wolf, R.M., & Klopfer, L.E. TSSA2. Chicago: University of Chicago: 1963.

5. Ryan, J.P. ABILEST. South Carolina Department of Education, 1983.

References

Ary, D., Jacobs, L.C., & Razavieh, A. Introduction to research in education (2nd ed.). New York: Holt, Rinehart, and Winston, 1979.

Benson, I.G. A comparison of three types of item analysis in test development using classical and latent trait methods (Doctoral dissertation, University of Florida, 1977). Dissertation Abstracts International, 1978, 39 (1-A), 243-244.

Benson, J., Crocker, & Wade. A comparison of three types of item analysis in test developing using classical and latent trait methods. Toronto, Canada: National Council on Measurement in Education, 1978. (ERIC Document No. ED 163 039).

Benson, J., & Wilson, M. A comparison of three types of test development procedures using classical and latent trait methods. San Francisco: American Educational Research in Education, 1979. (ERIC Document No. ED 178 556).

Bollenbacher, J. Testing in the public schools - A view from the middle. Educational Measurement - Issues and Practice, 1982, 1(1), 6-9; 25-26.

Brandt, R. On admissions testing: A conversation with Fred Hargadon. Educational Leadership, 1980, 37(8), 655-657.

Cook, L.L., & Hambleton, R.K. Application of latent trait models to the development of norm-referenced and criterion-referenced tests. Toronto, Canada: National Council on Measurement in Education, 1978. (ERIC Document No. ED 163 018).

Crehan, K.D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11(4), 255-262.

Denham, C.H. Criterion-referenced, domain-referenced, and norm-referenced measurement. Educational Technology, 1975, 15(12), 9-13.

Ebel, R.L. Essentials of educational measurement (3rd ed.). New Jersey: Prentice-Hall, 1979.

Goh, D.S. Empirical versus random item selection in the design of intelligence test short forms -- the WISC-R example. Applied Psychological Measurement, 1979, 3(1), 75-82.

Gronlund, N.E. Measurement and evaluation in teaching (4th ed.). New York: MacMillan, 1981.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Haney, W. Trouble over Testing. Educational Leadership, 1980, 37(8), 640-650.

Hatton, D.E. A comparison of latent-trait and classical item analysis in the development of achievement tests (Doctoral dissertation, University of Southern California, 1980). Dissertation Abstracts International, 1980, 41(4-A), 1551.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. Applied Statistics for the behavioral sciences. Chicago: Rand McNally, 1979.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14(6), 5-10.

Mead, R.J. Test Equating with the Dichotomous Rasch Model. SCALE Newsletter, 1982, 1(1), 1-4.

Millman, J. Passing scores and test lengths for domain referenced measures. Review of Educational Research, 1973, 43(2), 205-216.

Nairn, A., & Others. Class in the guise of merit. Educational Leadership, 1980, 37(8), 651-653.

Popham, W.J. Modern educational measurement. New Jersey: Prentice-Hall, 1981.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980. (Originally published, 1960).

Sax, G. Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, California: Wadsworth, 1981.

Smith, D. V. The effects of various item selection methods on the classification accuracy and classification consistency of criterion-referenced instruments. Toronto, Canada: American Educational Research Association, 1978. (ERIC Document No. ED 159 222).

Subkoviak, M.J., & Baker, F.B. Test Theory. In L.S. Shulman (Ed.), Review of research in education (Vol. 5). Itasca, Illinois: F.E. Peacock Publishers, 1978.

Wright, B.D., & Douglas, G.A. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 1977, 37(3), 573-586.

Wright, B.D., & Stone, M.H. Best test design. Chicago: Mesa Press, 1979.

Footnotes

[1]The Rasch item difficulty indexes and item-fit statistics were calculated on the responses of 343 students, reflecting the deletion of 16 students with excessive person-fit statistics and 12 students with perfect scores of 110. An additional 11 students with excessive person-fit statistics were identified in the second iteration of the BICAL-Version 3 (Mead, Wright, & Bell, Note 3) computer program. It is common practice to include the responses of these additional students in the estimation of all of the Rasch item statistics. However, there is no justification for the calculation of an ability estimate for any of these students. Consequently, the Rasch item statistics were calculated on a sample of 343 students, and the ability estimates were calculated for 332 students.

[2]As mentioned earlier, the BICAL-Version 3 (Mead, Wright, & Bell, Note 3) computer program was used to calibrate all of the items on the UT. The Rasch difficulties for these items can be found in Appendix C, and summary statistics for the difficulties on each test are shown in Table 5 (p. 36). An ANOVA on the mean difficulties for all four tests indicated statistically significant differences among them. Subsequent post-hoc analysis of all pairwise comparisons indicated significant differences for all comparisons with the exception of the UT and ST. These results indicated that the CT was the most difficult test, while the LTT was the easiest.

[3]The major reason for the reduction in the size of the sample can be attributed to the lack of ability estimates for those students who received zero or perfect scores on any of the four tests. In theory, the abilities associated with these scores cannot be calculated. In practice, the following procedure can be used to provide an ability estimate for either a zero or perfect score.

An initial ability estimate is identified as the ability of the closest score. This estimate is then adjusted by adding (or subtracting if appropriate) one, squared standard error associated with the initial ability estimate. For example, a score of 109 on the UT was associated with an ability of 5.26. This ability had a standard error of 1.01. Therefore, the ability estimate for a perfect score of 110 was estimated to be 6.28 (i.e., $5.26 + (1.01)^2$).

This procedure was used to estimate the abilities for the students who received zero or perfect scores on all of the tests. This resulted in a sample of 344 students, reflecting the deletion of only 27 students with misfitting response patterns.

The F-statistic associated with the repeated measures ANOVA on this data proved to be insignificant. Consequently, neither the procedure or its use were mentioned in the body of this paper.

Name_____

Add these numbers.

1 +1	9 +3	2 +1	+
9 +7	8 +8	9 +5	+
7 +4	5 +1	3 +2	+
7 +3	9 +9	1 +0	+
7 +1	3 +3	9 +0	

igit Addition and Subtraction Problems

s.

6 +6	8 +2	8 +4
8 +0	5 +0	4 +3
8 +3	0 +0	7 +7
6 +4	5 +3	3 +1

Name_____

Add these numbers.

4 +1	7 +2	5 +4	6 +5	7 +6	7 +0
9 +1	2 +0	8 +7	6 +0	4 +0	4 +2
6 +1	8 +6	5 +2	8 +5	6 +2	5 +5
4 +4	9 +4	9 +6	7 +5	3 +0	2 +2
9 +8	8 +1	6 +3	9 +2		

Appendix A (continued)

Name_____

Subtract these numbers.

8 -5	2 -0	6 -4	5 -3	3 -1	7 -1
3 -3	5 -5	9 -6	5 -2	7 -5	4 -4
9 -3	2 -2	9 -5	6 -3	9 -8	8 -1
8 -3	7 -0	1 -1	0 -0	9 -2	7 -6
2 -1	6 -6	5 -0			

Name_____

Subtract these numbers.

4 -3	8 -8	8 -2	7 -7	8 -7	9 -7
5 -4	6 -0	4 -0	6 -5	7 -4	5 -1
4 -1	3 -0	7 -2	9 -4	9 -0	8 -6
9 -9	9 -1	4 -2	6 -1	3 -2	7 -3
8 -4	6 -2	8 -0	1 -0		

Appendix B: Letter of Introduction

I would like to thank you in advance for allowing me to test the first grade students in your classroom. Having been a teacher myself, I am most appreciative of this opportunity.

This test is one of the major components in my dissertation study. I will be using the results of the test to examine several ways in which specific items can be described and selected for achievement tests. The study will not compare or evaluate the performance of the participating students, teachers, or schools. It is quite important that you realize that the test results will not be used for performance evaluation.

The test consists of 110 single-digit addition and subtraction problems. Although most first grade teachers have not completed their instruction in this area, it is important that the test be given at this time. I will examine the performance of each class and provide the teacher of that class with a summary of the results. This information may prove to be quite helpful to you in determining where future instruction can be best directed.

I would like you to begin testing the students on (date) and complete the testing by (date) . I have written the test in four sections - one page per section. I would like to have one page completed per day. I anticipate that each page will require about 15-20 minutes to complete. Please do not feel obliged to allow students more than 20 minutes if it is not convenient for you to do so.

Please return the tests to your principal when they are completed. I will pick them up and summarize the results for you as quickly as possible.

If you have any questions concerning the test, the summaries I will give you, or the study, I will be happy to answer them. I may be reached at 777-4887 or 359-2285.

Again, thanks very much for your help and cooperation.

Appendix C: Test Directions

You have been given a set of test booklets that are to be administered to your first grade students. Each test consists of four pages. Please administer one page per day to each student. Please begin testing around (date) so that you can finish around (date) .

On the first day, please give each child a test booklet. Ask them to write their initials in the appropriate space on the front page. When this has been completed, please read the following directions aloud:

> I would like to see how well you can do the math problems on this page. You will be able to do some of them very easily, but you may find some others very hard. If you want to use your fingers to count, you may do so. If you know the answer to a problem, write it under the problem. Try to do your very best, but don't get upset if you can't answer some of the problems. You will not be graded on this test.

If any of these points needs to be explained further, please feel free to do so.

After the students have finished the first page, collect the test booklets. DO NOT ALLOW THE STUDENTS TO WORK ON MORE THAN ONE PAGE PER DAY.

The directions for the second, third, and fourth days are the same as those for the first day. However, the students will answer page two on the second day, page three on the third day, and page four on the fourth day. Please make sure each student writes their initials at the top of each page, and please remember to collect the booklets at the end of each testing period.

Allow the students a reasonable amount of time to complete each page of the test. If it is convenient for you to permit even the slowest of students to finish, please do so. However, do not feel as though each student must finish the test.

When your students have completed all of the test pages, please return the test booklets to your principal.

Appendix D: Item Statistics
for the UT

Item	Classical		Rasch	
	Difficulty	Point-Biserial Discrimination	Difficulty	Item-Fit
A001	.97	.32	-2.96	3.30
A002	.86	.46	-0.62	0.14
A003	.95	.35	-2.37	3.24
A004	.82	.37	-0.04	4.96
A005	.90	.49	-1.08	0.53
A006	.80	.47	0.14	3.34
A007	.66	.52	1.33	2.29
A008	.68	.45	1.22	3.27
A009	.71	.56	0.88	-0.13
A010	.97	.43	-2.69	-0.56
A011	.96	.36	-2.69	0.33
A012	.87	.40	-0.73	2.34
A013	.76	.57	0.49	1.18
A014	.91	.42	-1.61	-0.74
A015	.91	.40	-1.49	1.98
A016	.77	.54	0.33	1.42
A017	.94	.23	-2.28	3.82
A018	.72	.48	0.81	2.60
A019	.82	.50	-0.17	1.78
A020	.62	.51	1.59	2.28
A021	.93	.45	-1.80	-0.80
A022	.80	.62	0.02	-0.13
A023	.85	.53	-0.58	0.74
A024	.91	.52	-1.55	-0.45
A025	.89	.57	-1.17	0.03
A026	.91	.50	-1.32	2.07
A027	.94	.45	-1.87	-0.18
A028	.96	.39	-2.58	-0.34
A029	.89	.46	-0.98	0.40
A030	.89	.47	-1.03	0.56
A031	.81	.42	-0.07	4.16
A032	.73	.37	0.81	5.04
A033	.97	.31	-3.28	0.79
A034	.90	.52	-1.27	-0.48
A035	.97	.41	-2.82	-0.54
A036	.69	.51	1.09	2.71
A037	.97	.36	-2.82	0.44
A038	.97	.33	-2.82	0.44
A039	.91	.49	-1.49	0.55

Appendix D (Continued)

Item	Classical		Rasch	
	Difficulty	Point-Biserial Discrimination	Difficulty	Item-Fit
A040	.91	.49	-1.55	1.59
A041	.69	.43	1.07	5.76
A042	.89	.54	-1.08	1.80
A043	.71	.48	0.84	3.24
A044	.83	.48	-0.26	5.13
A045	.88	.53	-1.03	2.23
A046	.88	.51	-0.73	-1.27
A047	.75	.53	0.59	2.50
A048	.66	.48	1.31	2.78
A049	.70	.49	0.93	4.47
A050	.93	.40	-1.95	1.35
A051	.92	.50	-1.55	-0.34
A052	.62	.52	1.55	1.69
A053	.85	.53	-0.65	2.46
A054	.78	.49	0.28	4.66
A055	.77	.58	0.19	1.66
S056	.83	.55	-0.01	-0.85
S057	.91	.34	-1.22	3.26
S058	.75	.62	0.67	-0.24
S059	.76	.67	0.57	0.43
S060	.74	.68	0.77	1.51
S061	.72	.65	0.98	1.73
S062	.73	.68	0.81	0.94
S063	.73	.64	0.91	0.78
S064	.70	.60	1.16	2.96
S065	.73	.67	0.91	0.66
S066	.70	.69	1.07	1.21
S067	.72	.64	0.95	1.24
S068	.64	.65	1.61	-0.16
S069	.70	.67	1.13	0.73
S070	.66	.64	1.41	1.01
S071	.72	.71	0.86	1.37
S072	.68	.71	1.22	2.01
S073	.69	.72	1.11	2.60
S074	.62	.65	1.71	1.11
S075	.88	.48	-0.81	0.67
S076	.61	.56	1.77	0.52
S077	.94	.36	-1.87	0.85
S078	.63	.67	1.63	2.47
S079	.67	.74	1.35	2.95
S080	.69	.68	1.18	1.67
S081	.70	.69	1.04	2.40
S082	.90	.38	-1.08	2.65
S083	.76	.70	0.59	1.89

Appendix D (Continued)

Item	Classical		Rasch	
	Difficulty	Point-Biserial Discrimination	Difficulty	Item-Fit
S084	.76	.70	0.62	2.02
S085	.74	.67	0.79	1.55
S086	.77	.65	0.52	0.66
S087	.70	.70	1.11	1.14
S088	.66	.66	1.39	1.05
S089	.71	.74	1.04	2.44
S090	.89	.43	-1.08	1.90
S091	.90	.41	-1.12	2.67
S092	.73	.72	0.81	1.72
S093	.66	.70	1.39	2.40
S094	.71	.74	1.00	2.86
S095	.72	.72	0.93	3.62
S096	.87	.48	-0.77	0.94
S097	.71	.73	1.04	3.25
S098	.65	.71	1.47	2.20
S099	.84	.45	-0.36	2.53
S100	.62	.70	1.67	1.13
S101	.73	.67	0.79	1.34
S102	.70	.75	1.07	2.41
S103	.70	.64	1.09	1.48
S104	.69	.78	1.11	3.30
S105	.67	.70	1.28	1.64
S106	.62	.69	1.67	1.97
S107	.68	.70	1.22	0.72
S108	.67	.75	1.28	2.85
S109	.86	.47	-0.58	2.13
S110	.83	.43	-0.33	2.38

Note: The "A" prefix refers to addition items, and the "S" prefix refers to subtraction items.

APPENDIX E: Raw Score and Rasch Ability Estimates for All Students

	Raw Score				Rasch Ability			
ID	UT	CT	LTT	ST	UT	CT	LTT	ST
101	84	26	23	24	.	.	.	.
102	102	28	30	29	3.08	3.60	.	3.79
103	101	30	26	28	2.95	.	1.77	3.06
104	109	29	30	30	5.26	4.31	.	.
105	85	24	26	25	1.64	2.38	1.77	1.99
106	56	19	14	15	.	.	.	.
107	108	30	30	30	4.55	.	.	.
108	93	27	25	25	2.18	3.17	1.46	1.99
109	101	28	29	28	2.95	3.60	3.38	3.06
110	46	0	15	10	-0.36	.	-0.73	-0.87
111	28	2	8	8	-1.41	-1.56	-2.15	-1.32
112	105	27	30	29	3.60	3.17	.	3.79
113	104	29	30	29	3.40	4.31	.	3.79
114	37	0	14	7	-0.86	.	-0.92	-1.57
115	104	28	29	29	3.40	3.60	3.38	3.79
116	109	30	29	30	5.26	.	3.38	.
117	79	29	19	22	.	.	.	.
118	108	30	29	30	4.55	.	3.38	.
119	102	28	29	26	3.08	3.60	3.38	2.27
120	90	28	28	23	1.96	3.60	2.63	1.53
121	96	26	28	26	2.42	2.86	2.63	2.27
122	95	29	25	28	2.34	4.31	1.46	3.06
201	95	26	28	28	2.34	2.86	2.63	3.06
202	97	24	27	27	2.52	2.38	2.14	2.61
203	46	3	17	10	-0.36	-1.13	-0.34	-0.87
204	107	30	30	29	4.14	.	.	3.79
205	92	19	26	25	2.10	1.56	1.77	1.99
206	103	25	30	26	3.23	2.60	.	2.27
207	63	3	21	15	0.49	-1.13	0.48	0.79
208	88	16	25	20	1.83	1.15	1.46	0.95
209	49	5	19	13	-0.21	-0.56	0.06	-0.28
210	97	25	28	26	2.52	2.60	2.63	2.27
211	87	18	28	24	1.76	1.42	2.63	1.75
212	99	25	28	29	2.72	2.60	2.63	3.79
213	57	2	22	12	0.20	-1.56	0.70	-0.47
214	36	8	14	11	-0.92	0.03	-0.92	-0.67
215	105	29	29	28	3.60	4.31	3.38	3.06
216	69	4	24	17	0.79	-0.81	1.18	0.43
217	100	28	28	28	2.83	3.60	2.63	3.06
301	22	0	11	6	-1.84	.	-1.51	-1.84
302	106	27	29	29	3.84	3.17	3.38	3.79
303	102	27	29	27	3.08	3.17	3.38	2.61
304	46	0	18	13	-0.36	.	-0.14	-0.28
305	103	29	30	29	3.23	4.31	.	3.79
306	83	20	25	24	1.52	1.70	1.46	1.75
307	90	26	28	24	1.96	2.86	2.63	1.75
308	88	13	28	23	1.83	0.76	2.63	1.53
309	101	26	27	29	2.95	2.86	2.14	3.79

Appendix E (continued)

	Raw Score				Rasch Ability			
ID	UT	CT	LTT	ST	UT	CT	LTT	ST
310	41	13	12	11	.	.	.	.
311	65	10	23	19	0.59	0.34	0.93	0.78
312	109	29	29	30	5.26	4.31	3.38	.
313	76	14	22	23	1.14	0.89	0.70	1.53
314	56	2	23	16	0.15	-1.56	0.93	0.25
315	62	3	20	17	0.44	-1.13	0.26	0.43
316	42	0	17	12	-0.58	.	-0.34	-0.47
317	57	0	21	15	0.20	.	0.48	0.79
318	101	28	30	29	2.95	3.60	.	3.79
319	58	1	21	16	0.24	-2.26	0.48	0.25
320	54	1	21	17	0.05	-2.26	0.48	0.43
401	105	29	29	29	3.60	4.31	3.38	3.79
402	108	29	29	28	4.55	4.31	3.38	3.06
403	77	24	19	21	1.19	2.38	0.06	1.14
404	89	20	23	25	1.89	1.70	0.93	1.99
405	103	28	30	29	3.23	3.60	.	3.79
406	103	28	30	30	3.23	3.60	.	.
407	84	16	25	23	1.58	1.15	1.46	1.53
408	98	28	26	25	2.61	3.60	1.77	1.99
409	108	29	30	30	4.55	4.31	.	.
410	108	30	30	30	4.55	.	.	.
411	107	29	30	28	4.14	4.31	.	3.06
412	98	28	27	28	2.61	3.60	2.14	3.06
413	103	25	29	24	3.23	2.60	3.38	1.75
414	107	28	29	30	4.14	3.60	3.38	.
415	85	20	24	24	1.64	1.70	1.18	1.75
416	65	13	22	19	0.59	0.76	0.70	0.78
417	30	3	11	10	.	.	.	.
418	93	21	25	27	2.18	1.85	1.46	2.61
419	107	29	29	29	4.14	4.31	3.38	3.79
420	82	18	24	25	1.47	1.42	1.18	1.99
501	102	27	27	28	3.08	3.17	2.14	3.06
501	61	7	21	15	0.39	-0.15	0.48	0.79
502	105	28	29	29	3.60	3.60	3.38	3.79
503	102	30	27	28	3.08	.	2.14	3.06
504	102	30	28	29	3.08	.	2.63	3.79
505	105	29	30	30	3.60	4.31	.	.
506	108	29	29	30	4.55	4.31	3.38	.
507	93	29	25	24	2.18	4.31	1.46	1.75
508	73	27	19	19	.	.	.	.
509	100	27	27	27	2.83	3.17	2.14	2.61
510	88	29	26	26	.	.	.	.
511	106	29	29	29	3.84	4.31	3.38	3.79
512	107	30	30	30	4.14	.	.	.

Appendix E (continued)

	Raw Score				Rasch Ability			
ID	UT	CT	LTT	ST	UT	CT	LTT	ST
513	89	27	23	26	1.89	3.17	0.93	2.27
514	92	28	27	25	2.10	3.60	2.14	1.99
515	97	26	25	25	2.52	2.86	1.46	1.99
516	108	30	30	30	4.55	.	.	.
517	108	30	30	30	4.55	.	.	.
518	100	25	30	29	2.83	2.60	.	3.79
519	109	30	30	30	5.26	.	.	.
520	99	30	28	27	2.72	.	2.63	2.61
601	86	12	27	21	1.70	0.62	2.14	1.14
602	104	28	28	28	3.40	3.60	2.63	3.06
603	101	29	29	28	2.95	4.31	3.38	3.06
604	106	29	29	30	3.84	4.31	3.38	.
605	105	30	28	27	3.60	.	2.63	2.61
606	107	28	29	28	4.14	3.60	3.38	3.06
607	70	29	21	19	.	.	.	.
608	110	30	30	30	.	.	.	.
609	108	30	30	30	4.55	.	.	.
610	107	30	29	29	4.14	.	3.38	3.79
611	105	30	28	28	3.60	.	2.63	3.06
612	105	29	29	29	3.60	4.31	3.38	3.79
613	101	27	27	27	2.95	3.17	2.14	2.61
614	102	28	29	28	3.08	3.60	3.38	3.06
615	102	28	28	29	3.08	3.60	2.63	3.79
616	105	26	30	28	3.60	2.86	.	3.06
617	107	29	30	28	4.14	4.31	.	3.06
618	98	27	29	26	2.61	3.17	3.38	2.27
619	106	28	30	29	3.84	3.60	.	3.79
620	99	24	28	27	2.72	2.38	2.63	2.61
621	98	23	27	29	2.61	2.19	2.14	3.79
701	53	0	20	15	0.00	.	0.26	0.79
702	107	30	29	30	4.14	.	3.38	.
703	97	30	28	29	2.52	.	2.63	3.79
704	56	12	21	15	0.15	0.62	0.48	0.79
705	105	28	30	29	3.60	3.60	.	3.79
706	85	22	23	25	1.64	2.02	0.93	1.99
707	45	0	15	12	-0.42	.	-0.73	-0.47
708	104	29	26	28	3.40	4.31	1.77	3.06
709	47	6	17	12	.	.	.	.
710	58	4	24	16	0.24	-0.81	1.18	0.25
711	49	3	14	13	-0.21	-1.13	-0.92	-0.28
712	101	29	28	29	2.95	4.31	2.63	3.79
713	87	17	27	23	1.76	1.29	2.14	1.53
714	90	23	26	24	1.96	2.19	1.77	1.75
715	88	24	23	23	1.83	2.38	0.93	1.53

Appendix E (continued)

ID	Raw Score				Rasch Ability			
	UT	CT	LTT	ST	UT	CT	LTT	ST
716	103	29	29	28	3.23	4.31	3.38	3.06
717	82	21	22	21	1.47	1.85	0.70	1.14
718	105	30	28	29	3.60	.	2.63	3.79
719	109	30	30	30	5.26	.	.	.
720	67	6	20	20	0.69	-0.34	0.26	0.95
801	35	2	13	8	-0.97	-1.56	-1.12	-1.32
802	59	4	21	15	0.29	-0.81	0.48	0.79
803	38	0	15	11	-0.80	.	-0.73	-0.67
804	57	20	15	20	.	.	.	.
805	72	7	24	21	0.94	-0.15	1.18	1.14
806	11	2	6	2	-2.89	-1.56	-2.63	-3.40
807	46	0	20	11	-0.36	.	0.26	-0.67
808	51	5	16	15	-0.10	-0.56	-0.53	0.79
809	92	24	29	23	2.10	2.38	3.38	1.53
810	80	19	19	22	.	.	.	.
811	101	30	29	29	2.95	.	3.38	3.79
812	76	20	18	24	.	.	.	.
813	108	30	30	30	4.55	.	.	.
814	77	23	19	26	.	.	.	.
815	104	30	29	29	3.40	.	3.38	3.79
816	59	3	20	13	0.29	-1.13	0.26	-0.28
817	63	2	23	18	0.49	-1.56	0.93	0.60
818	96	29	28	25	2.42	4.31	2.63	1.99
819	15	4	9	4	-2.45	-0.81	-1.93	-2.48
820	103	29	28	30	3.23	4.31	2.63	.
821	69	20	18	21	.	.	.	.
822	61	4	20	18	0.39	-0.81	0.26	0.60
823	103	30	29	29	3.23	.	3.38	3.79
901	54	3	21	11	0.05	-1.13	0.48	-0.67
902	51	0	14	14	-0.10	.	-0.92	-0.10
903	80	10	29	21	1.35	0.34	3.38	1.14
904	69	4	25	18	0.79	-0.81	1.46	0.60
905	19	1	8	6	-2.09	-2.26	-2.15	-1.84
906	94	19	28	27	2.26	1.56	2.63	2.61
907	94	22	27	28	2.26	2.02	2.14	3.06
908	80	22	21	23	.	.	.	.
909	67	3	22	17	0.69	-1.13	0.70	0.43
910	50	2	18	11	-0.15	-1.56	-0.14	-0.67
911	63	0	23	17	0.49	.	0.93	0.43
912	54	0	23	12	0.05	.	0.93	-0.47
913	95	24	29	24	2.34	2.38	3.38	1.75
914	105	29	28	30	3.60	4.31	2.63	.
915	84	21	27	23	1.58	1.85	2.14	1.53
916	96	24	27	27	2.42	2.38	2.14	2.61

Appendix E (continued)

	Raw Score				Rasch Ability			
ID	UT	CT	LTT	ST	UT	CT	LTT	ST
917	90	21	24	21	1.96	1.85	1.18	1.14
918	58	7	18	15	0.24	-0.15	-0.14	0.79
919	106	28	29	28	3.84	3.60	3.38	3.06
920	57	1	22	16	0.20	-2.26	0.70	0.25
921	109	29	30	30	5.26	4.31	.	.
922	98	28	27	30	2.61	3.60	2.14	.
923	96	24	30	27	2.42	2.38	.	2.61
1001	110	30	30	30	.	.	.	.
1002	108	30	30	29	4.55	.	.	3.79
1003	13	2	2	3	-2.66	-1.56	-4.06	-2.88
1004	102	30	28	28	3.08	.	2.63	3.06
1005	87	21	24	20	1.76	1.85	1.18	0.95
1006	103	28	28	29	3.23	3.60	2.63	3.79
1007	25	0	13	6	-1.62	.	-1.12	-1.84
1008	54	3	21	14	0.05	-1.13	0.48	-0.10
1009	109	30	30	30	5.26	.	.	.
1010	58	3	20	16	0.24	-1.13	0.26	0.25
1011	79	13	26	20	1.30	0.76	1.77	0.95
1012	60	1	21	14	0.34	-2.26	0.48	-0.10
1013	42	1	17	10	-0.58	-2.26	-0.34	-0.87
1014	74	9	23	20	1.04	0.19	0.93	0.95
1015	23	4	8	7	.	.	.	.
1016	100	30	29	27	2.83	.	3.38	2.61
1017	46	5	17	10	-0.36	-0.56	-0.34	-0.87
1018	50	3	18	16	-0.15	-1.13	-0.14	0.25
1019	60	0	21	15	0.34	.	0.48	0.79
1020	87	26	26	25	1.76	2.86	1.77	1.99
1101	106	29	30	28	3.84	4.31	.	3.06
1102	102	27	29	29	3.08	3.17	3.38	3.79
1103	108	30	29	29	4.55	.	3.38	3.79
1104	110	30	30	30	.	.	.	.
1105	104	27	29	28	3.40	3.17	3.38	3.06
1106	110	30	30	30	.	.	.	.
1107	110	30	30	30	.	.	.	.
1108	108	30	28	30	4.55	.	2.63	.
1109	110	30	30	30	.	.	.	.
1110	101	24	29	30	2.95	2.38	3.38	.
1111	101	27	29	27	2.95	3.17	3.38	2.61
1112	110	30	30	30	.	.	.	.
1113	106	28	30	27	3.84	3.60	.	2.61
1114	107	29	29	30	4.14	4.31	3.38	.
1115	110	30	30	30	.	.	.	.
1116	108	30	30	30	4.55	.	.	.
1117	104	28	29	28	3.40	3.60	3.38	3.06

Appendix E (continued)

ID	Raw Score				Rasch Ability			
	UT	CT	LTT	ST	UT	CT	LTT	ST
1118	104	29	29	29	3.40	4.31	3.38	3.79
1119	96	23	30	25	2.42	2.19	.	1.99
1120	104	30	30	29	3.40	.	.	3.79
1121	88	24	25	27	1.83	2.38	1.46	2.61
1122	87	30	23	25	.	.	.	.
1123	104	27	30	27	3.40	3.17	.	2.61
1201	48	12	19	11	-0.26	0.62	0.06	-0.67
1202	83	25	24	25	1.52	2.60	1.18	1.99
1203	87	23	25	24	1.76	2.19	1.46	1.75
1204	106	29	27	30	3.84	4.31	2.14	.
1205	85	28	27	20	1.64	3.60	2.14	0.95
1206	102	25	28	26	3.08	2.60	2.63	2.27
1207	110	30	30	30	.	.	.	.
1208	99	27	28	29	2.72	3.17	2.63	3.79
1209	83	23	24	24	1.52	2.19	1.18	1.75
1210	97	28	26	27	2.52	3.60	1.77	2.61
1211	71	18	20	21	0.88	1.42	0.26	1.14
1212	86	16	29	21	1.70	1.15	3.38	1.14
1213	95	28	28	27	2.34	3.60	2.63	2.61
1214	106	28	30	28	3.84	3.60	.	3.06
1215	97	25	27	26	2.52	2.60	2.14	2.27
1216	98	29	29	29	2.61	4.31	3.38	3.79
1217	41	11	13	9	.	.	.	.
1218	105	30	29	28	3.60	.	3.38	3.06
1219	101	27	27	28	2.95	3.17	2.14	3.06
1220	98	28	29	29	2.61	3.60	3.38	3.79
1221	75	21	21	20	.	.	.	.
1222	3	0	1	1	-4.45	.	-4.82	-4.19
1223	90	20	28	26	1.96	1.70	2.63	2.27
1224	107	29	28	30	4.14	4.31	2.63	.
1301	90	26	26	25	1.96	2.86	1.77	1.99
1302	96	25	25	27	2.42	2.60	1.46	2.61
1303	104	28	30	28	3.40	3.60	.	3.06
1304	108	29	30	30	4.55	4.31	.	.
1305	109	29	30	30	5.26	4.31	.	.
1306	106	30	28	29	3.84	.	2.63	3.79
1307	102	29	29	30	3.08	4.31	3.38	.
1308	92	20	26	23	2.10	1.70	1.77	1.53
1309	103	28	27	29	3.23	3.60	2.14	3.79
1310	89	28	24	25	.	.	.	.
1311	100	26	26	29	2.83	2.86	1.77	3.79
1312	108	28	30	29	4.55	3.60	.	3.79
1313	103	30	30	29	3.23	.	.	3.79
1314	109	30	29	30	5.26	.	3.38	.

Appendix E (continued)

ID	Raw Score				Rasch Ability			
	UT	CT	LTT	ST	UT	CT	LTT	ST
1315	104	30	29	30	3.40	.	3.38	.
1316	67	7	21	21	0.69	-0.15	0.48	1.14
1317	102	29	28	28	3.08	4.31	2.63	3.06
1318	104	27	26	29	3.40	3.17	1.77	3.79
1319	104	29	29	29	3.40	4.31	3.38	3.79
1320	102	27	28	28	3.08	3.17	2.63	3.06
1321	103	30	28	30	3.23	.	2.63	.
1401	103	25	30	29	3.23	2.60	.	3.79
1402	106	29	29	28	3.84	4.31	3.38	3.06
1403	63	5	18	16	0.49	-0.56	-0.14	0.25
1404	83	12	27	17	1.52	0.62	2.14	0.43
1405	97	28	29	27	2.52	3.60	3.38	2.61
1406	61	1	22	16	0.39	-2.26	0.70	0.25
1407	103	30	27	28	3.23	.	2.14	3.06
1408	104	29	29	29	3.40	4.31	3.38	3.79
1409	71	9	24	20	0.88	0.19	1.18	0.95
1410	66	4	25	20	0.64	-0.81	1.46	0.95
1411	86	21	24	24	1.70	1.85	1.18	1.75
1412	104	25	30	29	3.40	2.60	.	3.79
1413	89	17	29	23	1.89	1.29	3.38	1.53
1414	96	27	28	27	2.42	3.17	2.63	2.61
1415	110	30	30	30	.	.	.	.
1416	110	30	30	30	.	.	.	.
1417	85	23	23	25	.	.	.	.
1418	109	30	29	29	5.26	.	3.38	3.79
1419	107	30	30	29	4.14	.	.	3.79
1420	79	16	26	22	1.30	1.15	1.77	1.33
1501	81	16	28	22	1.41	1.15	2.63	1.33
1502	88	19	27	27	1.83	1.56	2.14	2.61
1503	104	28	29	28	3.40	3.60	3.38	3.06
1504	103	27	28	26	3.23	3.17	2.63	2.27
1505	108	29	29	30	4.55	4.31	3.38	.
1506	77	17	22	22	.	.	.	.
1507	37	0	16	10	-0.86	.	-0.53	-0.87
1508	53	0	19	12	0.00	.	0.06	-0.47
1509	50	0	20	12	-0.15	.	0.26	-0.47
1510	48	0	18	11	-0.26	.	-0.14	-0.67
1511	44	1	19	10	-0.47	-2.26	0.06	-0.87
1512	32	3	14	10	-1.16	-1.13	-0.92	-0.87
1513	57	11	22	12	.	.	.	.
1514	69	17	20	17	.	.	.	.
1601	109	30	30	30	5.26	.	.	.
1602	108	30	30	30	4.55	.	.	.
1603	100	28	29	25	2.83	3.60	3.38	1.99

Appendix E (continued)

ID	Raw Score				Rasch Ability			
	UT	CT	LTT	ST	UT	CT	LTT	ST
1604	101	27	28	30	2.95	3.17	2.63	.
1605	107	29	30	30	4.14	4.31	.	.
1606	95	25	26	26	2.34	2.60	1.77	2.27
1607	102	29	29	27	3.08	4.31	3.38	2.61
1608	99	27	25	26	2.72	3.17	1.46	2.27
1609	109	29	30	30	5.26	4.31	.	.
1610	108	30	30	30	4.55	.	.	.
1611	99	26	30	28	2.72	2.86	.	3.06
1612	98	24	27	24	2.61	2.38	2.14	1.75
1613	92	23	25	28	2.10	2.19	1.46	3.06
1614	35	9	12	12	.	.	.	.
1615	33	2	15	9	-1.09	-1.56	-0.73	-1.09
1616	13	0	6	4	-2.66	.	-2.63	-2.48
1701	20	5	6	7	.	.	.	.
1702	51	2	19	12	-0.10	-1.56	0.06	-0.47
1703	102	30	26	29	3.08	.	1.77	3.79
1704	93	30	27	26	2.18	.	2.14	2.27
1705	99	26	29	26	2.72	2.86	3.38	2.27
1706	105	30	30	29	3.60	.	.	3.79
1707	108	29	30	28	4.55	4.31	.	3.06
1708	68	19	22	17	0.74	1.56	0.70	0.43
1709	46	5	18	10	-0.36	-0.56	-0.14	-0.87
1710	110	30	30	30	.	.	.	.
1711	98	29	27	26	2.61	4.31	2.14	2.27
1712	106	28	30	28	3.84	3.60	.	3.06
1713	103	29	28	27	3.23	4.31	2.63	2.61
1714	108	29	30	29	4.55	4.31	.	3.79
1715	106	29	30	30	3.84	4.31	.	.
1716	106	28	30	28	3.84	3.60	.	3.06
1717	103	30	30	27	3.23	.	.	2.61
1718	104	29	28	29	3.40	4.31	2.63	3.79
1719	108	30	30	30	4.55	.	.	.
1801	95	25	26	24	2.34	2.60	1.77	1.75
1802	98	26	27	29	2.61	2.86	2.14	3.79
1803	95	23	30	25	2.34	2.19	.	1.9 1
1804	105	29	29	29	3.60	4.31	3.38	3.79
1805	105	30	30	29	3.60	.	.	3.79
1806	95	28	28	25	2.34	3.60	2.63	1.99
1807	100	28	27	29	2.83	3.60	2.14	3.79
1808	100	28	30	29	2.83	3.60	.	3.79
1809	43	6	14	13	.	.	.	.
1810	78	16	25	22	1.25	1.15	1.46	1.33
1811	106	29	30	29	3.84	4.31	.	3.79
1812	62	8	18	19	0.44	0.03	-0.14	0.78

Appendix E (continued)

	Raw Score				Rasch Ability			
ID	UT	CT	LTT	ST	UT	CT	LTT	ST
1813	84	17	26	23	1.58	1.29	1.77	1.53
1814	92	28	27	27	2.10	3.60	2.14	2.61
1815	104	29	30	29	3.40	4.31	.	3.79
1816	86	24	27	25	1.70	2.38	2.14	1.99
1817	46	0	18	14	-0.36	.	-0.14	-0.10
1901	50	5	19	16	-0.15	-0.56	0.06	0.25
1902	99	30	27	30	2.72	.	2.14	.
1903	106	29	30	28	3.84	4.31	.	3.06
1904	109	29	30	30	5.26	4.31	.	.
1905	101	26	26	25	2.95	2.86	1.77	1.99
1906	48	8	16	12	-0.26	0.03	-0.53	-0.47
1907	49	5	20	14	-0.21	-0.56	0.26	-0.10
1908	108	29	30	29	4.55	4.31	.	3.79
1909	97	27	29	26	2.52	3.17	3.38	2.27
1910	91	21	29	27	2.03	1.85	3.38	2.61

Note: A "." indicates an undetermined ability estimate.

Appendix F: Pertinent Information from the Secondary Analysis of the Data

Table 1

Descriptive Statistics for All Groups on All Tests

Group	N	UT Mean (S.D.)	CT Mean (S.D.)	LTT Mean (S.D.)	ST Mean (S.D.)
1	93	.45 (.14)	.16 (.19)	.58 (.17)	.44 (.15)
2	185	.86 (.08)	.83 (.15)	.90 (.09)	.87 (.09)
3	93	.977 (.01)	.976 (.03)	.985 (.02)	.979 (.03)

Appendix F (Continued)

Table 3

Repeated Measures ANOVA Data Comparing
All Tests - Group 2

Source	SS	df	MS	F	PROB.
Between Tests	.38	3	.13	24.56	< .001
Test * Subject Interaction	2.87	552	.01		

Tukey Pairwise Comparisons for Group 2

$$_{95}Q_{4,368} = 3.63$$

$$Q_{ut\text{-}ct} = 4.91$$

$$Q_{ut\text{-}ltt} = -7.06$$

$$Q_{ut\text{-}st} = -2.03$$

$$Q_{ct\text{-}ltt} = -11.97$$

$$Q_{ct\text{-}st} = -6.94$$

$$Q_{ltt\text{-}st} = 5.02$$

Appendix F (Continued)

Table 2

Repeated Measures ANOVA Comparing
All Tests - Group 1

Source	SS	df	MS	F	PROB.
Between Tests	8.83	3	2.94	244.21	< .001
Test * Subject Interaction	3.33	276	.01		

Tukey Pairwise Comparisons for Group 1

$$.95Q_{4,368} = 3.63$$

$$Q_{ut-ct} = 25.37$$

$$Q_{ut-ltt} = -11.84$$

$$Q_{ut-st} = .92$$

$$Q_{ct-ltt} = -37.21$$

$$Q_{ct-st} = -24.45$$

$$Q_{ltt-st} = 12.75$$

Appendix F (Continued)

Table 4

Repeated Measures ANOVA Comparing
All Tests - Group 3

Source	SS	df	MS	F
Between Tests	.004	3	.001	3.47*
Test * Subject Interaction	.11	276	.0004	

Tukey Pairwise Comparisons

$$_{95}Q_{4,368} = 3.63$$

$$Q_{ut\text{-}ct} = .07$$

$$Q_{ut\text{-}ltt} = -3.85$$

$$Q_{ut\text{-}st} = -.59$$

$$Q_{ct\text{-}ltt} = -3.92$$

$$Q_{ct\text{-}st} = -.66$$

$$Q_{ltt\text{-}st} = 3.26$$